December 6, 2011

THE COMPUTATIONAL MAGIC OF THE VENTRAL STREAM

Online archived report: usage instructions. This is version 2 of a report first published online on July 20, 2011 (npre.2011.6117.1) and it is still preliminary. I expect there will be more versions in the near future because the work is promising and ongoing. This version does not yet contain all the appendices because some of them are just project ideas for students in my group. They will appear in due time. I marked some of the statements that suggest a research project depending on whether it is a safe, significant project ($\bigodot$) or a more blue-sky, risky project ($\bigoplus$). I do not expect that this paper will ever be published in the usual journals. "Success" for a paper published in this way would consist, I believe, of making an impact – measured in terms of citations for instance – and perhaps of being eventually "reviewed" in sites such as Wired or Slashdot or Facebook or even in a News and Views-type article in traditional journals like Science or Nature.

Tomaso Poggio*,† (sections 4 and 5.0.1 with Jim Mutch*; section 1.1.1 with Joel Leibo* and appendix 7.6 with Lorenzo Rosasco†)

$\star$ *CBCL, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA*

† Istituto Italiano di Tecnologia, Genova, Italy

ABSTRACT. I argue that the sample complexity of (biological, feedforward) object recognition is mostly due to geometric image transformations and conjecture that a main goal of the feedforward path in the ventral stream – from V1, V2, V4 and to IT – is to learn-and-discount image transformations.

In the first part of the paper I describe a class of simple and biologically plausible memory-based modules that learn transformations from unsupervised visual experience. The main theorems show that these modules provide (for every object) a *signature* which is invariant to local affine transformations and approximately invariant for other transformations. I also prove that, in a broad class of hierarchical architectures, signatures remain invariant from layer to layer. The identification of these memory-based modules with complex (and simple) cells in visual areas leads to a theory of invariant recognition for the ventral stream.

In the second part, I outline a theory about hierarchical architectures that can learn invariance to transformations. I show that the memory complexity of learning affine transformations is drastically reduced in a hierarchical architecture that factorizes transformations in terms of the subgroup of translations and the subgroups of rotations and scalings. I then show how translations may be automatically selected as the only learnable transformations during development by enforcing small apertures – eg small receptive fields – in the first layer.

In a third part I show that the transformations represented in each area can be optimized in terms of storage and robustness, as a consequence determining the tuning of the neurons in the area, rather independently (under normal conditions) of the statistics of natural images. I describe a model of learning that can be proved to have this property, linking in an elegant way the spectral properties of the signatures with the tuning of receptive fields in different areas.

A surprising implication of these theoretical results is that the computational goals and some 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. In particular, simple and complex cells do not directly care about oriented bars: their tuning is a side effect of their role in translation invariance. Across the whole ventral stream the preferred features reported for neurons in different areas are only a symptom of the invariances computed and represented.

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

- invariance to small translations is the main operation of V1;
- invariance to larger translations and small local scalings and rotations is the main characteristic of 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;
- each cell's tuning properties are shaped by visual experience of image transformations during developmental and adult plasticity;
- the type of transformations that are learned from visual experience depend on the size of the receptive fields and thus on the area (layer in the models) – assuming that the size increases with layers;
- the mix of transformations 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 (eg face tuned cells);
- features must be discriminative and invariant: invariance to specific transformations is the primary determinant of the tuning of cortical neurons rather than statistics of natural images.
- homeostatic control of synaptic weights during development is required for hebbian synapses.
- 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 theory is broadly consistent with the current version of HMAX. It explains it and extends it in terms of unsupervised learning, a broader class of transformation invariance and higher level modules. The goal of this paper is to sketch a comprehesive theory with little regard for mathematical niceties. If the theory turns out to be useful there will be scope for deep mathematics, ranging from group representation tools to wavelet theory to dynamics of learning.

CONTENTS

1. 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 [25–28] and references therein, see also section 7.1), which are faithful to the physiology and 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 the empirical and the modeling advances the ventral stream is still a puzzle: until now we do not have 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 indeed some promise or whether it is nonsense.

As background to this paper, I assume that the content of past work of my group on models of the ventral stream is known from old papers [25–28] to more recent technical reports [13–17]. See also the section *Background* in Supp. Mat. [21]. After writing most of this paper I found a few interesting and old references about transformations, invariances and receptive fields, see [8,11,19]. I stress that a key assumption of this paper is that in this initial theory and modeling i can neglect subcortical structures such as the pulvinar as well as cortical backprojections.

1.1. **Recognition is difficult because of image transformations.**

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, if the images of objects are rectified with respect to all transformations.

1.1.1. *Empirical Evidence (with J. Leibo).* 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 Figure 1). The figure shows that if we factor out all transformations in images of many different dogs and many different horses – obtaining "normalized" 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 normalized 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).

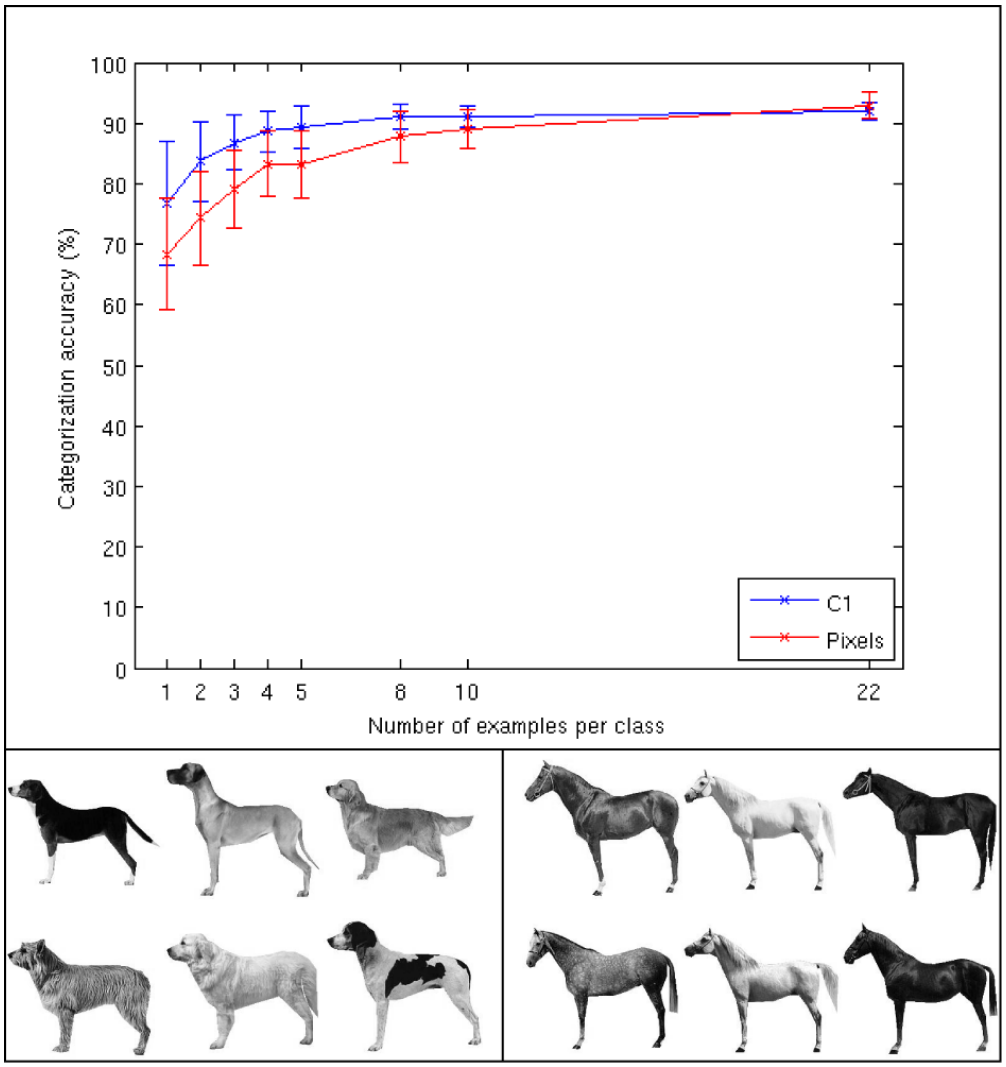


FIGURE 1. *Images of dogs and horses, 'normalized' with respect to image transformations. A regularized least squares classifier (linear kernel) tested on more than 150 dogs and 150 horses does well with little training. Error bars represent +/- 1 standard deviation computed over 100 train/test splits.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

1.1.2. *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 I 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. It has been known for a long time that this problem can be solved under the assumption that correspondence of enough points between stored models and a new image can be computed. As one of the simplest such results, it turns out that under the assumption of correspondence, two training images are enough for orthographic projection (see [32]). Recent techniques for normalizing for affine transformations are now well developed (see [34] for a review). Various attempts at learning transformations have been reported over the years by Rao and Hinton among others [6, 23]. See for additional references the paper by Hinton [6].

The goal here is instead to explore approaches to the problem that do not rely on explicit correspondence operations and provide a plausible theory for the ventral stream.

In summary, my conjecture is that *the main goal of the ventral stream is to learn to factor out image transformations*. I plan to show that from this conjecture may consequences follow 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.

1.2. **Plan of the paper.** In the introduction I described the conjecture that the sample complexity of object recognition is mostly due to geometric image transformations, eg different viewpoints , and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. The first part of section 2 deals with theoretical results that are rather independent of specific models; they are the main results of this paper. They are motivated by layered architectures "looking" at images, or at "neural images" in the layer below, through a number of small "apertures" corresponding to receptive fields, on a 2D lattice. I have in mind a *memory-based architecture* in which learning consists of "storing" (the main argument is developed for a "batch" version but a more plausible "online" version is possible) patches of neural activation. The main results are

(1) recording transformed templates - *the templatebook* – provides a simple and biologically plausible way to obtain an invariant signature for any new object, which can be used for recognition. This is the *invariance lemma* in section 2.2.
(2) several *aggregation* (eg pooling) functions including the energy function and the the max preserve invariance of signatures in a hierarchical architecture. This is the *aggregation theorem* of section 2.3.

Section 3 shows how the natural factorization of the affine group in subgroups implies that, wrt memory complexity, hierarchies are significantly superior to nonhierarchical architectures. I also outline a preliminary theory of how different types of invariances may be learned at the bottom and at the top of the hierarchy, depending on the sizes of the receptive fields. In particular, I discuss two topics:

(1) most importantly, the transformation "learned" at a layer depends on the aperture size; this is the *stratification theorem*.

(2) less importantly, global transformations can be approximatd by local affine transformations (the *approximation lemma*)

Section 4 discusses ideas of how transformations learned in the way described in the first section may determine tuning properties of neurons in different layers of the hierarchy. In particular, I show that the spatiotemporal spectral properties of the templatebook depend on the transformations (represented through stored examples). The connection with tuning of cells is provided by a *linking theorem* stating that plausible forms of associative Hebbian learning connect the spectral properties of the templatebook to the tuning of simple cells at each layer.

Together with the arguments of the previous sections this theory-in-fieri provides the following highly 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 which transformations are learned during development and then factored out during normal processing; that class-specific transformations are learned and represented at the top of the hierarchy; and that the transformation represented in an area influences the tuning of the neurons in the area. The final section puts everything together in terms of a class of models which extends HMAX.

1.3. **Remarks.**

- **Generic and class-specific transformations** We distinguish (as I did in past papers, see [22,25]) between generic image-based transformations that apply to every object, such as scale and translation, and class specific transformations, such as rotation in depth, that can apply (not exactly) to a 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. This predicts modularity of the architecture for recognition because of the need to route – or reroute – information to transformation modules which are class specific [14].
- **Memory-based architectures, correlation and associative learning** The architecture assumed in this paper can be regarded as a case of memory-based learning of transformations by storing 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 (correlational and Hebbian) learning is the main mechanism. The key is a Foldiak-type rule: *cells that fire together are wired together*. At the level of C cells this rule determines *classes of equivalence* between simple cells reflecting observed *time correlations in the real world, that is transformations* of the image. The main function of the hierarchy is thus to learn different types of invariances via association of templates memorized during transformations in time. There is a general and powerful principle here, induced by the markovian (eg differential equations) physics of the world, that allows associative labeling of stimuli based on their temporal contiguity. We may call this principle *The principle of Probable Time Smoothness*$^{\oplus}$.
- **Spectral theory and receptive fields** The third part of the paper discusses the possibility of a *spectral theory* trying to link specific transformations and invariances to tuning properties of cells in each area through Hebbian learning rules. If this research program bears fruit then we will have a rather complete theory. Its most surprising implication

would be 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. In fact one may – in the extreme – argue that a Foldiak-type rule determines by itself the hierarchical organization of the ventral stream, the transformations that are learned and the receptive fields in each visual area.

- **Subcortical structures and recognition** I am neglecting 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$^{\oplus}$.

2. THEORY: MEMORY-BASED INVARIANCE

In this section I have in mind a hierarchical layered architecture as shown in Figure 2. I also have in mind a computational architecture that is memory-based in the sense that it stores sensory inputs and does very little in terms of additional computations: it computes normalized dot products and max-like aggregation functions. However, 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 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 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 and all their stored transformations is computed. A vector called "signature" is produced by an aggregation function such as a max over the dot products with each template and its transformations. Suppose now that the same image is shown again but in this case transformed. 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 the max (on the group), acting on the signature, will then be invariant to the learned transformation.

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, actually entire analytic functions in $\mathbb{R}^2$ since they are bandlimited by the optics of the eye to about $60\ cycles/degree$. 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 outer cones (which touch each other) which is 27 seconds of arc. The sampled image is then processed by retinal neurons; the result is communicated to the LGN and then primary visual cortex through the optic nerve, consisting of axons of the retinal ganglion cells. At the LGN level there are probably several neural "images": they may

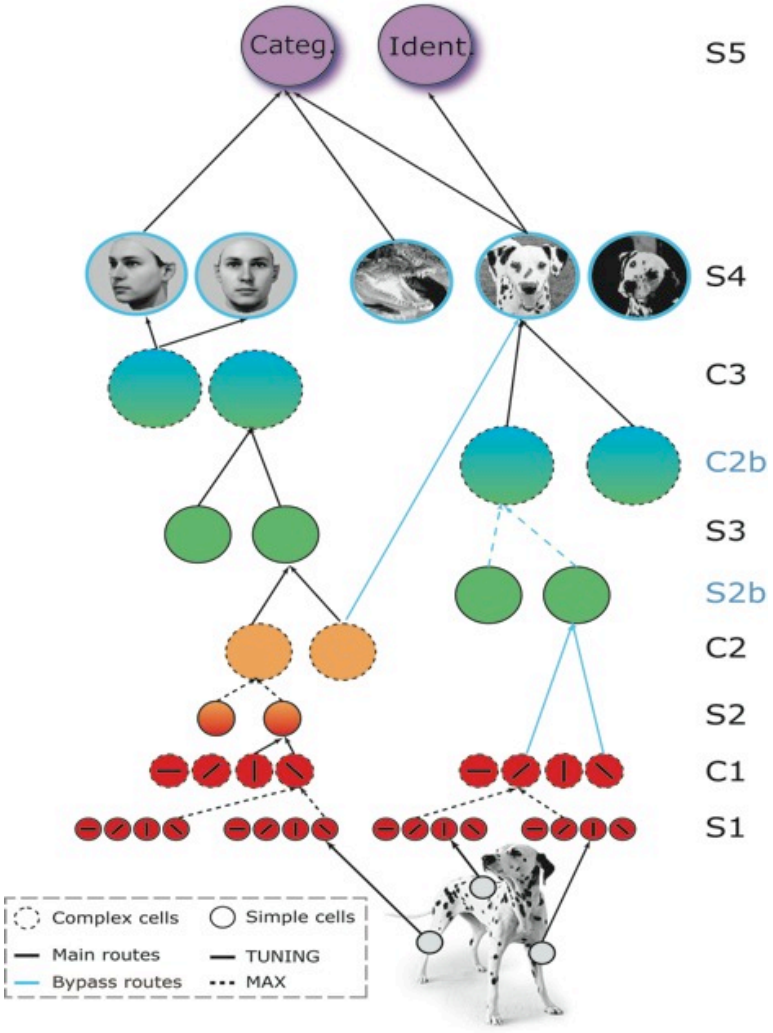


FIGURE 2. *Hierarchical feedforward model of the ventral stream – a modern interpretation of the Hubel and Wiesel proposal (see [24]). The theoretical framework proposed in this paper provides foundations for this model and how the synaptic weights may be learned during development (and with adult plasticity). It also suggests extensions of the model adding class specific modules at the top.*

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 different scales. There is also high-pass filtering in time at the level of the retina which can be approximated as a zero-th order approximation by a time derivative or more accurately as a filter providing, in the Fourier domain, $\beta F(\omega_x, \omega_y, \omega_t) + i\omega_t F(\omega_x, \omega_y, \omega_t)$. Thus the neural image seen by the cortex is bandpass in space and time. The finest grain of it is set by the highest spatial frequency (notice that if λ_u corresponds to the highest spatial frequency then sampling at the Shannon rate, eg on a lattice with edges $\frac{\lambda_u}{2}$ preserves all the information.)

2.2. **Templatebooks and Invariant Signatures.**

2.2.1. *Signatures of images and Johnson-Lindenstrauss.* As humans we are estimated to be able to recognize on the order of 50K object classes through single images, each one with a dimensionality of 1M pixels (or ganglion cell axons in the optic nerve). This means high discrimination requirements.

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 2 is in terms of an overcomplete set of measurements on the image, a vector that we will call here a *signature*.

We assume here that the *nature of the measurements is not 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. A more mathematical *motivation* is provided by a theorem due to Johnson and Lindenstrauss. Their classic result is (informal version):

Theorem 1. *(Johnson-Lindenstrauss) 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.*

This means that each of the n d-dimensional points can be represented in terms of n k-dimensional points, via k random projections. For example the vector f can be represented as the k-dimensional vector resulting by projecting f on the range of the k, d matrix R

$$Rf = \begin{pmatrix} \tau_{1,1} & \tau_{1,2} & \cdots & \tau_{1,d} \\ \tau_{2,1} & \tau_{2,2} & \cdots & \tau_{2,d} \\ \cdots & \cdots & \cdots & \cdots \\ \tau_{k,1} & \tau_{k,2} & \cdots & \tau_{k,d} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_d \end{pmatrix} \tag{1}$$

The theorem – and related results – suggests that since there are no special conditions on the projections (in fact they are typically assumed to be random) most measurements will work, 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).

I describe here a set of measurements and a process of using them – both consistent with HMAX (a generalization of it) and with the known physiology and psychophysics.

FIGURE 3. *Number of intersection per line (out of an arbitrary but fixed set) provides a highly efficient signature for OCR.*

FIGURE 4. *See previous figure.*

Informally, the components of the signature vector are the *normalized dot products* of the image (or image patch) f w.r.t. a set of *templates* $\tau_{\mathbf{i}}$, $i = 1, \cdot, k$, which are image patches themselves. More formally, I define *templatesets*:

2.2.2. *Templatebooks.* We start defining a set of templates which can be random images, that is *A "templateset" is a set of neural "image" patches* $\tau_{\mathbf{i}}, \mathbf{i} = \mathbf{1}, \cdots, \mathbf{k}$.

We can think of the templateset $\mathbb{T}_{set}$ as a vector of images

$$\mathbb{T}_{set} = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \cdots \\ \tau_k \end{pmatrix}. \tag{2}$$

Consider now the templateset $\mathbb{T}_{set}$ as a column vector of images – thus a tensor since each element is an image – which can be thought of as k templates (eg different patches of an image, corresponding for instance to different parts of the same object), learned or observed at the same time $t = 1$:

Assume now that the image of the object undergoes a geometric transformation due to motion (of the eye or of the object). For simplicity we will assume that the transformation can be well approximated locally (see later) by an affine transformation in R^2. We are really interested in perspective projection of 3D transformations of rigid objects and more generally nonrigid transformations (such as a face changing expression or rotating in depth; or a body changing pose) but we consider here mostly affine transformations in $\mathbb{R}^2$. We will denote as $\tau_1^{g^1}, \tau_1^{g^2}, \cdots, \tau_1^{g^N}$ the images of the same patch at frames (i.e. instants in time) $t = 1, \cdots, N$. Thus each element of the templateset $\mathbb{T}_{set}$ is replaced by all the transformations of it to provide the templatebook ($\mathbb{T}$ for simplicity)

$$\mathbb{T} = \begin{pmatrix} \tau_1^{g^0} \\ \tau_1^{g^1} \\ \tau_1^{g^2} \\ \cdots \\ \tau_1^{g^N} \\ \tau_2^{g^0} \\ \tau_2^{g^1} \\ \tau_2^{g^2} \\ \cdots \\ \tau_2^{g^N} \\ \cdots \\ \tau_d^{g^0} \\ \tau_d^{g^1} \\ \tau_d^{g^2} \\ \cdots \\ \tau_d^{g^N} \end{pmatrix}. \tag{3}$$

We call the matrix $\mathbb{T}$ a *templatebook generated by a templateset and a set (possibly a finite group) of transformations* g^i.

Each subset of row of the template book $\mathbb{T}$, such as

$$\begin{pmatrix} \tau_i^{g^0} \\ \tau_i^{g^1} \\ \tau_i^{g^2} \\ \cdots \\ \tau_i^{g^N} \end{pmatrix}. \tag{4}$$

can be thought of as the set of simple cells, see Figure 23, that are pooled by the same complex cell. Each simple cell can be thought of as corresponding to a frame in a video associated with a complex cell[1]. The formal definition is

Definition 1. *A templatebook consists of templateset* $\tau_{\mathbf{i}}$ *and all its transforms obtained by the action on* $\tau_{\mathbf{i}}$ *of elements* g^i *of a set of transformations.*

2.2.3. *Signatures.* We now define the signature of an image – a set of features which represents the fingerprint of the image – used for classification.

Definition 2. *The expanded "signature" of* $\mathbf{f}$ *wrt the templatebook* $\mathbb{T}$ *is the vector* $\Sigma_f = K(f, \tau_i)$ *for* $i = 1, \cdots, d$*, where* K *is the normalized kernel*

[1]Later in section 5 we will see a slightly different interpretation, in terms of principal components of the templatebook.

$$K(g,h) = \frac{g \circ h}{(|g \circ g||h \circ h|)^{1/2}}.$$

Thus

$$Sigma_f = \begin{pmatrix} f \circ \tau_1^{g^0} \\ f \circ \tau_1^{g^1} \\ f \circ \tau_1^{g^2} \\ \cdots \\ f \circ \tau_1^{g^N} \\ f \circ \tau_2^{g^0} \\ f \circ \tau_2^{g^1} \\ f \circ \tau_2^{g^2} \\ \cdots \\ f \circ \tau_2^{g^N} \\ \cdots \\ f \circ \tau_d^{g^0} \\ f \circ \tau_d^{g^1} \\ f \circ \tau_d^{g^2} \\ \cdots \\ f \circ \tau_d^{g^N} \end{pmatrix}. \tag{5}$$

Notice that the components of the signature vector can be seen as the projections in the Johnson-Lindestrauss theorem (nit random anymore!). As we will see later, I am thinking of a recursive computation of signatures, from layer to layer from small patches to large ones.

2.2.4. *Image transformations.* Images are functions on $\mathbb{R}^2$. A geometric transformation T on an image $f(x,y)$ is defined as $Tf\mathbf{x} = f(T\mathbf{x})$. For instance, $Tf(x,y) = f(u,v)$, with $(u,v) = T(x,y)$. 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 $\mathbb{R}^2$. It can be written as a semidirect product: $Aff(2,\mathbb{R}) = GL(2,\mathbb{R}) \times \mathbb{R}^2$ where $GL(2,\mathbb{R})$ acts on $\mathbb{R}^2$ in the natural manner. Later we will assume that transformations induced by all the elements of G – or a subgroup of G – are contained in the templatebook (for instance "all" the translations of a patch $\tau_{1,1}$ may be represented (in practice this will be within resolution and range constraints)).

We say that a group G acts on a space X, if $\forall g \in G$ is a mapping $Tg : XX$ such that if $g2g1 = g3$, then $Tg1(Tg2(x)) = Tg3(x)\ \forall x \in X$. We say that X X is a *homogeneous space of* G if fixing any $x_0 \in X$, the set $Tg(x_0)$ ranges over the whole of X as g ranges over G.

2.2.5. *The invariance lemma.* Consider the expanded signature vector Σ_f corresponding to an image patch f with respect to a templateset, as defined earlier, and a set of transformations. Thus $\Sigma_f = (f \cdot \tau_1, f \cdot \tau_2, \cdots, f \cdot \tau_n)$, where

$$f \cdot t_i = \frac{\int f(x,y)\tau_i(x,y)dxdy}{[(\int f(x,y)dxdy)(\int \tau_i(x,y)dxdy)]^{1/2}} \tag{6}$$

Consider now geometric transformations $Tf(x,y) = f(u,v)$. We call the transformation *uniform* if the Jacobian $J(x,y) = constant$. As a major example T may correspond to affine transformations on the plane eg $\mathbf{x}' = A\mathbf{x} + \mathbf{t}_\mathbf{x}$ with A a nonsingular matrix.

Then the following *invariance lemma* holds

Lemma 1. *The expanded signature Σ_f of f with respect to the templateset $\mathbb{T}_{set}$ is equal to the expanded signature Σ_{gf} of gf w.r.t. the templateset $g\mathbb{T}_{set}$ for uniform transformations.*

Proof sketch (see Appendix for proof): It is enough to consider the effect on one of the coordinates.

$$f \cdot \tau_i = \frac{\int f(x,y)\tau_i(x,y)dxdy}{[(\int f(x,y)dxdy)(\int \tau_i(x,y)dxdy)]^{1/2}} = \tag{7}$$

$$= \frac{\int f(u,v)\tau_i(u,v)|J(u,v)|dudv}{[(\int f(u,v)|J(u,v)|dudv)(\int \tau_i(u,v)|J(u,v)|dudv)]^{1/2}} =$$

$$= \frac{\int f(u,v)\tau_i(u,v)dudv}{[(\int f(u,v)dudv)(\int \tau_i(u,v)dudv)]^{1/2}} = Tf \cdot T\tau_i$$

□

The invariance lemma implies that independently of the templates – and how selective they are – the signature they provide can be completely invariant to a geometric transformation which is uniform over the pooling region. We will see later an architecture for which signatures are invariant recursively through layers. The actual templates themselves do not enter the argument: the set of similarities of the input image to the templates need not be high in order to be invariant.

Remark: The analysis of relevant nongeometric transformations *is an interesting open question for future research*$^{\odot}$.

2.2.6. *Closed Templatebooks and Invariance from One Training Example.* Let us call a templatebook *closed*[2] relative to a set of transformations $g \in G$, if for any transformation g and for any entry $\tau_{i,j}$ there is a k s.t. $g\tau_{i,j} = \tau_{i,k}$. Then an object that has image f can be recognized from a single example, *independently* of the unknown transformation. For now just notice that if there is a single training image f of an object and that if a closed templatebook is available in memory, then the signatures of f are identical (apart from order) to the signatures of gf (see Appendix in section 7.2).

[2]The more standard notation is to assume $\tau \in \mathcal{H}$ with $\mathcal{H}$ a Hilbert space and to assume that there is for every $g \in G$ a unitary operator $U(g) : \mathcal{H} \to \mathcal{H}$ such that U is a unitary representation of G. Then for $\tau \in \mathcal{H}$ the *orbit* of τ is the set of vectors that can be reached by the action of the representation: thus

$$orbit\ of\ \tau = U(g)\tau \quad s.t.\ g \in G.$$

A subspace $\mathcal{B} \in \mathcal{H}$ is called invariant if it contains all its orbits. To avoid confusing the reader we use in the main text the term "closed" instead of the term "invariant". Notice that G may be the whole affine group or a subgroup such as the translation group.

2.3. **Invariant aggregation functions.** For each complex cell there is a set of templates – that is a template and its transformations over which the complex cell is pooling. Let us use each element in the matrix below as the vector of measurements relative to the transformations of one object-patch, listed in the row in the order in which they appear during the transformations. So the column index effectively runs through time and through the simple cells pooled by the same complex cell:

$$\Sigma_f^{\mathbb{T}} = \begin{pmatrix} f \circ \tau_{1,2} & f \circ \tau_{1,2} & \cdots & f \circ \tau_{1,N} \\ f \circ \tau_{2,1} & f \circ \tau_{2,2} & \cdots & f \circ \tau_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ f \circ \tau_{M,1} & f \circ \tau_{M,2} & \cdots & f \circ \tau_{M,N} \end{pmatrix}.$$

Definition 3. *A signaturebook $\Sigma_f^{\mathbb{T}}$ holds the signatures of f for a set of templates $\tau_{1,\cdot}, \cdots, \tau_{D,\cdot}$ wrt a set of transformations g that belong to a group G.*

Now one needs to aggregate each set of simple cells into a number, so that it can provides to a higher layer a signature – a vector of m components (as many components as complex cells).

To accomplish this, the model (for instance HMAX) uses an *aggregation function* $\biguplus$ such as a max of $f \circ \tau_{i,j}$ over j or the average of $f \circ \tau_{i,j}$ over j or the average of $(f \circ \tau_{i,j})^2$ over j. The latter operation is an approximation of a sigmoidal function describing a threshold operation of a neuron or of a dendritic spike. These aggregation operations can be approximated by the generalized polynomial

$$y = \frac{\sum_{i=1}^{n} w_i \, {x_i}^p}{k + \left(\sum_{i=1}^{n} {x_i}^q \right)^r} \tag{8}$$

for appropriate values of the parameters (see [12]). 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 $energy \quad operation(x) = ||x||_2$.

We need to make sure that the aggregation function is indeed invariant. We can prove this using the invariance lemma 1 in the following way.

Assume the following *Aggregation learning rule*:

Assume that templatebooks have been acquired during development at layer i. The rule for assigning the signature Σ_f of a new image f at layer $i+1$ is to select for each row j of $\Sigma_f^{\mathbb{T}}$ (eg for each complex cell) the value $f \circ \tau_i^* = \biguplus_j f \circ \tau_{i,j}$.

The current version of HMAX works computes $f \circ \tau_i^* = max_j f \circ \tau_{i,j}$. We will focus later on the *energy operation* $f \circ \tau_i^* = \frac{1}{n}(\sum_1^n (f \circ \tau_{i,j})^2)^{\frac{1}{2}}$.

We have the following *aggregation theorem*:

Theorem 2. *If the set of templates is closed under actions of a group then the max of the signature of f is invariant to transformations of f, that is $\Sigma_f = \Sigma_{gf}$ for $g \in G$ (an example of G is $G = Aff(2, \mathbb{R})$.*

Proof sketch in the case of $\biguplus = max$ (see Appendix for proof): By assumption the aggregation function chooses for an element of the signature at the higher level the max over templates t (see rule above), providing as component of the higher level signature $f \circ t^*$. After this choice

is made during learning, assume that the new image to recognize at run time is $f' = Tf$. We claim that $max_{t\in T} f' \circ t = f \circ t^*$. Assume the opposite eg $max_{t\in T} f' \circ t \geq f \circ t^*$. This implies that $max_{t\in T} f \circ t \geq f \circ t^*$, which contradicts the assumption. $\square$
This means that at runtime the max will (in the noiseless situation) provide the same value – independently of g. Across complex cells this says that the *signature is invariant wrt G from layer to layer*.

Note that the same invariance result can be proved for other aggregation functions such as the average of powers of the elements of the row of Σ_f . The result holds for groups (and not only the affine group).

The approach taken with the aggregation function (above) is an example of averaging over the group to obtain invariants, eg $R_G[f(x)] = \frac{1}{|G|}\sum_{g\in G} f(g(x))$ and is based on the group property (and the possibility to factorize a group such as the affine group in subgroups).

All of the above justifies the following definition

Definition 4. *The "G-invariant signature" of* **f** *is the vector* $\Sigma^G(f) = \biguplus_j f \circ K(f, \tau_{i,j})$ *where* K *is the normalized kernel and* $\tau_{i,j} = g\tau_{i,1}$ *for some* $g \in G$*, where* G *is a group and the templates are closed wrt* G*.* $\Sigma^G(f)$ *is invariant to any transformation* gf *of* f*,* $\quad g \in G$*.*

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. In fact

Theorem 3. *For subgroups of the affine group on* $\mathbb{R}^2$ *by averaging over the subgroup (ie* $R_G[f(x)] = \frac{1}{|G|}\sum_{g\in G} f(g(x))$ *) the following aggregation function on a patch* $I(x)$ *gives a number which is invariant to any transformation of* I*:* $R_G[I(x)] = \frac{1}{|G|}\sum_{g\in G} |\int I(x) g \circ \chi(x) dx|^2$ *where* $\chi(x)$ *are characters of the group.*

In general, *signatures at various levels may be used by a classifier*.

Remarks: *Aggregation functions: invariance and discriminability.*

- It is to be expected that different aggregation functions, all invariant, have different power of discriminability and noise robustness. For instance, the arithmetic average will tend to make signatures that are invariant but 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 [24] to choose the max over the average[3]. In any case, *this is an interesting open question for future research*$^{\oplus}$.
- 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.
- Consider a group of transformations such as the similitude group of the plane SIM(2) consisting of translations, dilations and rotations. Its structure corresponds to a semidirect product structure $(SO2 \times \mathbb{R}) \times \mathbb{R}^2$. Rotations and dilations commute and the subgroup of translations is invariant. An aggregation function such as the energy operation

[3]Notice that zero-crossings – which are relatives of maxima – of one-octave bandpass functions contain all the information about the function.

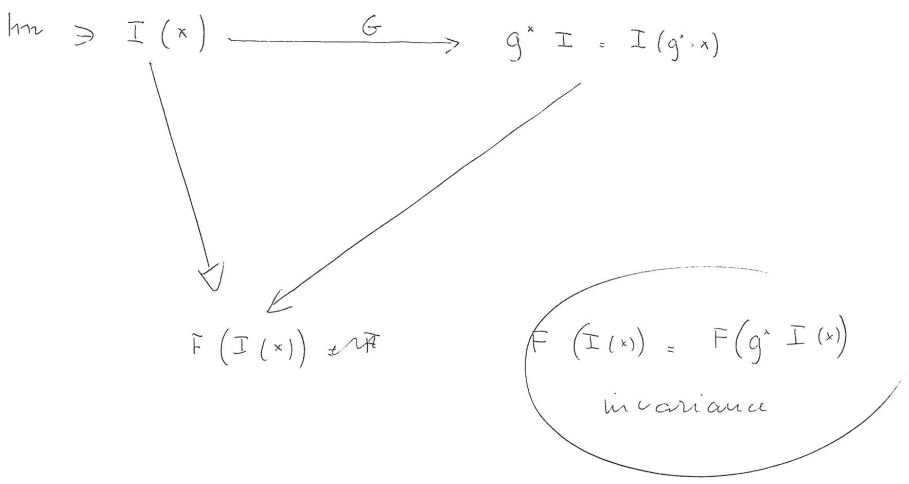


FIGURE 5. *An invariant function F (under the group G). Suggested by F. Anselmi.*

or the max can be defined over the whole group or as composition of max over subgroups (corresponding to a hierarchical computation).

3. THEORY: HIERARCHY OF INVARIANCES

3.1. **Factorization of Invariances and Hierarchies.** I assume that the goal of the ventral stream is to be invariant to transformations of images in the sense described in the previous section, by storing at least one *image τ and all its transformations.*

The transformations I consider include 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 section 2 can deal – exactly or approximately – with all these transformations.

For simplicity of the analysis of this section let me consider here transformations described by $Aff(2, \mathbb{R})$–the affine group on the plane. A representation of the affine group can be given

in terms of the matrices

$$\mathbf{g} = \begin{pmatrix} a & b & t_x \\ d & e & t_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{9}$$

where the g are representations of the affine group $Aff(2,\mathbb{R})$ which is an extension of $GL(2,\mathbb{R})$ by the group of translations in $\mathbb{R}^2$. It can be factorized as a semidirect product: $Aff(2,\mathbb{R}) = GL(2,\mathbb{R}) \times \mathbb{R}^2$ where $GL(2,\mathbb{R})$ acts on $\mathbb{R}^2$ in the natural manner, that is

$$\mathbf{G} = \begin{pmatrix} a & b & t_x \\ d & e & t_y \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a & b & 0 \\ d & e & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{10}$$

where the matrix

$$\mathbf{L} = \begin{pmatrix} a & b & 0 \\ d & e & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{11}$$

can be itself decomposed in different ways for instance by decomposing the 2×2 matrix

$$\mathbf{L}' = \begin{pmatrix} a & b \\ d & e \end{pmatrix} \tag{12}$$

as $L' = U\Sigma V^T$, where U and V are orthogonal, eg rotations, and Σ is diagonal, eg scaling. Notice that Equation 10 is the standard representation of a "forward" affine transformation (as used in graphics) in which translation follows rotation and scaling (which commute). Our setup requires the inverse transformation and thus the inverse order (in this example). Notice also that while the affine group is not abelian, each of the subgroups is abelian, under mild conditions.

The *key* observation is that a one layer architecture with a single large aperture looking at the whole image would need to store the images generated from an initial τ by acting on it with all the rotations, scalings, translations *and all their combinations*. Under the assumptions of the "back-of-the-envelope" estimate reported in the introduction, this may require on the order of $10^8 - 10^{14}$ distinguishable images. Compare this to a hierarchy that stores in the first layer all translations, in the second all rotations, in the third all scalings. I denote the sequence of transformations as $\mathfrak{S} \circ \mathfrak{R} \circ \mathfrak{T}$, eg the application first of translation, then rotation and finally scaling (to undo the inverse forward transformation of a pattern). Using the memory-based module of the previous section, I need to store 3 templates and their transformations in each of 3 layers: τ_1 will be stored with all its translations in layer 1, τ_2 will be stored with all its rotations in layer 2 and τ_3 will be stored with all its scalings in layer 3. The total number of stored templates will roughly be $(10^3 + 10^3 + 10) - (10^5 + 10^5 + 10^2)$. This gives on the order of $10^3 - 10^5$ templates that need to be stored in all the layers – a significant saving even for a memory-based system like the brain. Thus we have a *factorization lemma*:

Lemma 2. *Suppose that the matrix representation of a group of transformations G can be factorized as $G_1 \cdot G_2$. Suppose that the storage complexity of the memory-based module for G is $n = n_1 \cdot n_2$. The storage complexity of a hierarchy of two layers of memory-based modules is $n_1 + n_2$.*

Proof: see Appendix$\odot$

It is cute to think that evolution may have been tempted by such an memory complexity advantage of a hierarchical architecture for discounting transformations. Evolution, however, has a problem. How can it program development of a visual system in such a way that the system is selectively and sequentially exposed only to translations during development, followed at some later time by rotations and scalings?

A possible answer is provided by the following observation. Suppose that the first layer consists of an array of "small apertures" – in fact corresponding to the receptive fields of V1 cells – and focus on one of the apertures. I will show that the only transformations that can be estimated by a small aperture are small translations, even if the transformation of the image is more complex.

3.2. **Stratification Theorem (with Mahadevan).** Consider, given two or more frames observed through a small aperture of size r, the task of estimating an affine transformation. I conjecture that the following statement holds:

Stratification Theorem *Assume pointwise correspondence between successive frames. Assume a fixed amount of measurement noise. For increasing aperture size (relative to resolution of the image) the first transformation that can be estimated reliably above a fixed threshold, is translation. For large aperture size, more complex transformations can be estimated or approximated, such as the image transformation induced by perspective projection of rotation of a face in 3D. In general, the transformations that can be estimated for small apertures are generic and independent of the specific object, whereas some the transformations for large apertures are class specific.*

Proof sketch (see Appendix for proof$\odot$):

We define an aperture to be small when it is measured in terms of spatial frequencies of the input "images" (which can be neural "images" provided by the previous layer), see section 2.1: thus an aperture which is on the order of $10 \times \lambda_u$ is small. Informally the proof is based on the observation that estimation of translation requires estimating two numbers with condition number equal to 1. Rotation requires estimating three numbers (pure rotation plus translation) with condition number equal to 1. Scaling requires estimating four numbers (asymmetric scaling plus translation) – and the condition number may be bad. Therefore translation requires the least amount of bits, followed by rotation, asymmetric scaling and full affine. Thus increasing aperture size corresponds to increasing number of bits (which is correct in terms of the memory-based module). In summary

- one corresponding point over two frames is enough for estimating the two translation parameters
- one additional point is needed to estimate the rotation angle or the scaling
- three points are needed to estimate the 6 affine parameters
- more points require a larger aperture for a fixed amount of noise in the measurements
- estimation of rotation and scale requires first estimation of the center of rotation and the focus of expansion, respectively. It is natural, therefore, to estimate translation first.

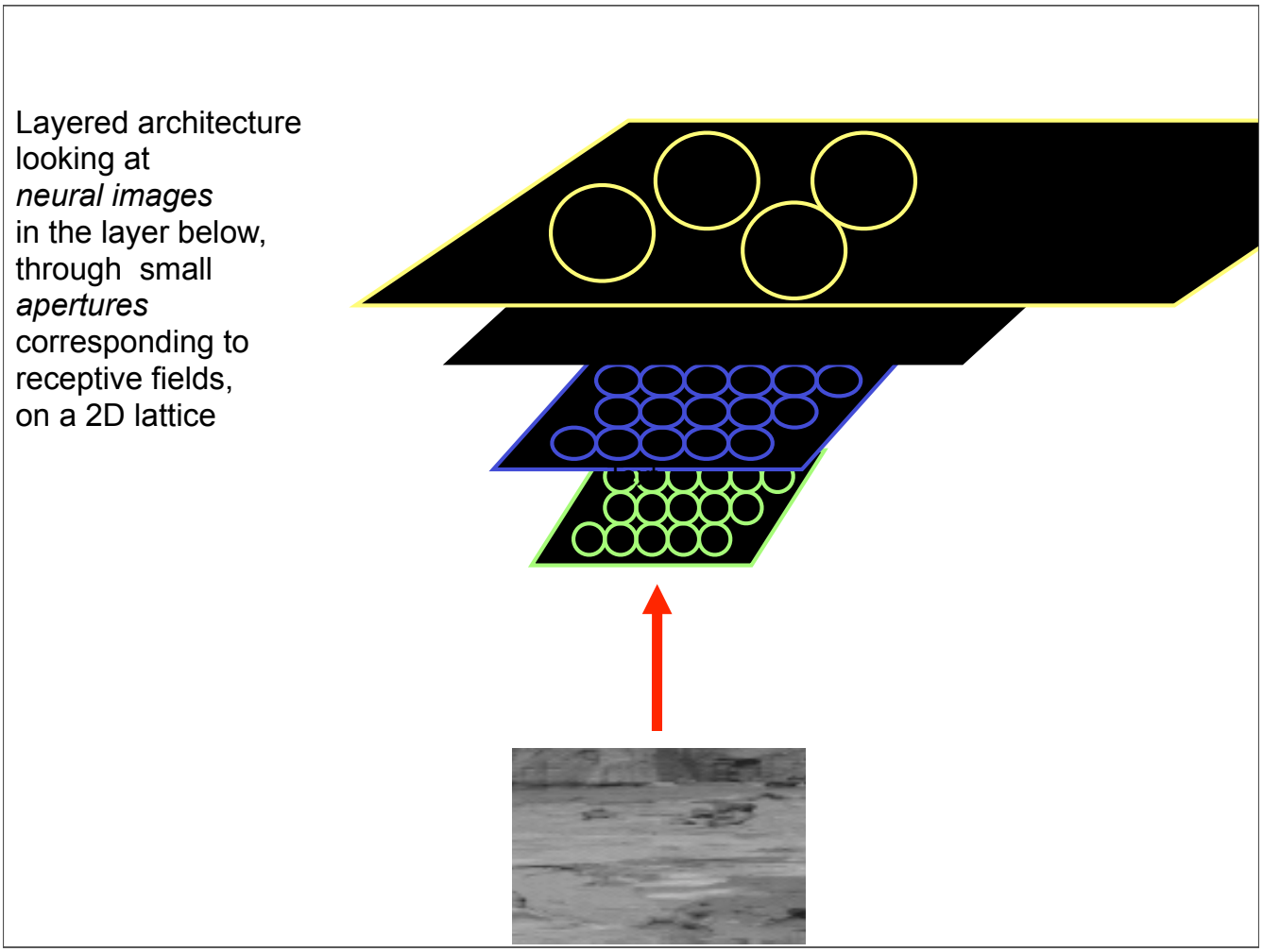


FIGURE 6. *A layered architecture with apertures of increasing size (in reality overlapping).*

Thus the last transformation that can be estimated when the size of an aperture goes to zero is translation. □

Notice that for large apertures a transformation that is locally affine or that can be approximated by local affine transformations may not be globally affine. An example is the transformation induced by rotation in depth of a face.

Notice that

- the aperture size limits the size of translations that can be "seen" and estimated from a single aperture (eg a single receptive field or cell)
- there is a tradeoff between large aperture for more robust estimation of larger transformations and the correspondence problem which increases in difficulty (I expect this argument is equivalent to using correlation between the two frames to estimate parameters)

- growing aperture size corresponds to increasingly complex features (from 1-point features to 3-point features)
- for affine transformations the condition number of the problem is 1 for isotropic scaling; it can grow arbitrarily large depending on the asymmetry of the scaling.

For my argument, I need only to establish that the only transformations that are represented in a first layer of small apertures, eg "learned" and later discounted (at run time) are (small) translations. As I will discuss later, the output of an aggregation operation on the first layer templatebooks will be invariant to most of the small translations[4]. Repeating the process at a second and higher layers with the same memory-based mechanism and max gives invariance to larger translations and more complex transformations. It is unlikely$^{\oplus}$ that there exists a strict order of learned transformations (after translations in the first layer), such as uniform scaling followed by rotations followed by nonuniform scaling. In part this is because scaling and rotations commute. Notice however that the argument of the factorization lemma still holds.

By using this property of small apertures, evolution can solve the problem of how to enforce "presentation" of translations only in the first layer during development and thus enforce factorization. The fact that our reasoning can focus on affine transformations is related to a related property of architectures with local apertures, as shown in the next section.

3.2.1. *The local affine approximation lemma.* Geometric transformations of the image can be locally approximated by their "linear" term, eg affine transformations in $\mathbb{R}^2$ (see section 7.3). It seems possible to approximate any global transformation of the image arbitrarily well for patch size going to zero (and increasing number of patches) if appropriate affine transformations are used for each patch. It seems that the following *Approximation Lemma* should hold:

Lemma 3. *Local affine transformations $Af\, f(2, \mathbb{R})$ on image patches can approximate any smooth transformation in $\mathbb{R}^2$ within a given tolerance with a sufficient number of patches.*

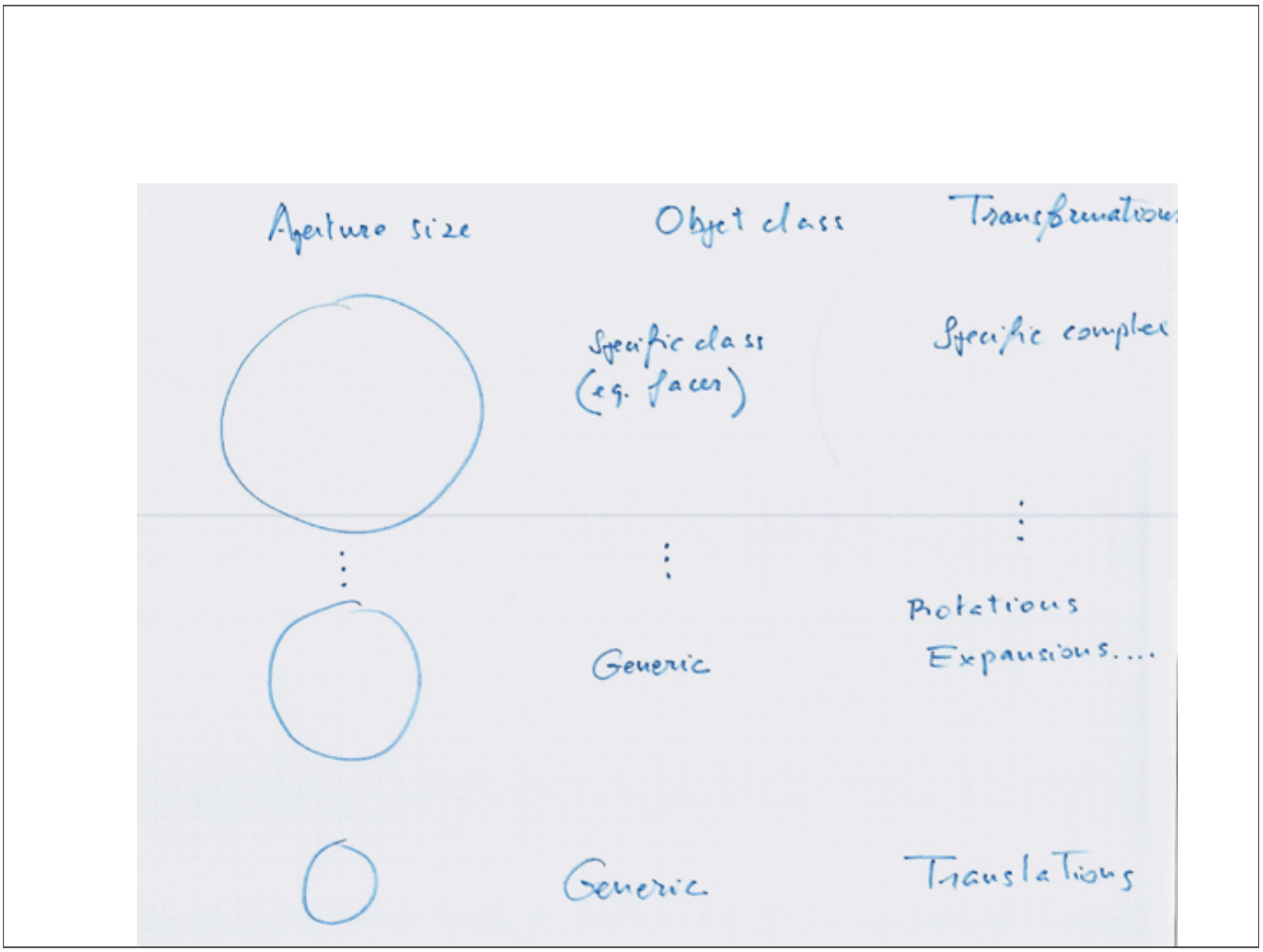
Proof sketch (see Appendix for proof$^{\odot}$): If $\mathbf{x}_0$ is a point at the center of the patch and the transformation $Tf(\mathbf{x}) = f(T\mathbf{x})$ and T is differentiable at $\mathbf{x}_0$ then its derivative is given by $J_T(\mathbf{x}_0)$. In this case, the linear map described by $J_T(\mathbf{x}_0)$ is a linear approximation of T near the point $\mathbf{x}_0$, in the sense that

$$T(\mathbf{x}) = T(\mathbf{x}_0) + J_T(\mathbf{x}) \cdot (\mathbf{x} - \mathbf{x}_0) + o(||\mathbf{x} - \mathbf{x}_0||). \tag{13}$$

□

The affine approximation lemma suggests that under some conditions of biological significance *a set of local "receptive fields" on a lattice may be appropriate for representing (and discounting) small global non uniform deformations of the image.* The conjecture may be another justification for the biological evolution of architectures consisting of many local apertures, as a first layer. For remarks on this chapter see section 7.4. It is important to stress that transformations which are not affine will be approximated by local affine transformations but the signature will not be completely invariant.

[4]Signatures at the output of the first layer may not be completely invariant because of "boundary effects" at each aperture in the present implementation of the model



Sunday, August 7, 2011

FIGURE 7. *The conjecture is that receptive field sizes determines the type of transformation that can be learned. In particular, small apertures allow learning of (small) translations only. As the receptive field size increases the learned transformations are increasingly complex and less generic that is specific for specific classes of objects – from translations in V1 to rotations in depth of faces in patches of IT.*

3.3. **Transformations: Stratification and Peeling Off.** In the previous section 2, I gave some results and observations on how images could be recognized from a single training image in an invariant way by learning implicitly a set of transformations in terms of a templatebook. Let me now summarize the main points of section 3. A one-layer system comprising the full image (a large aperture) would require that a memory-based module store all the transformations induced by all elements g of the full group of transformations. It is however more efficient memory-wise to follow a factorization approach in terms of the subgroups. This corresponds to a hierarchical architecture dealing with translations first, followed by layers dealing with other transformations. Luckily for evolution, a first layer of small apertures "sees" only translations – and can then store the associated transformed templates. The reason for a *hierarchy of modules*

dealing each with specific subgroups of transformations is that the memory complexity becomes additive instead of multiplicative (the case of a non-hierarchical architecture). Thus it is natural that layers with apertures of increasing size learn and discount transformations – in a sequence, from simple and local transformations to complex and less local ones. Learning transformations during development in a sequence of layers corresponds to the term *stratification*, while the sequential use of the layers at run time, one after the other, corresponds to the term *peeling off*. In the previous subsection 3.2 I conjecture that stratification appears because the kind of transformations that are learned depend on the aperture (eg the receptive field size of a complex cell), which increases from V1 to IT. In other words I conjecture that receptive field size determines type of transformations learned and their sequence. In the final layers the structure and the spectral properties of the templatebooks may depend on the natural statistics of image transformations. Ongoing simulations (J. Mutch) should find out how many layers are needed to learn the full affine transformations over a range of translations. Notice that small apertures can only learn and represent small transformations, since larger translations will usually yield little correlation between subsequent frames.

3.3.1. *Class-specific transformations.* Consider a hierarchy of layers that have learned invariance to most affine transformations of generic patterns, especially translation and scales. An example of such an invariant system is the (hardwired) HMAX model at the C2 level. Suppose that a non-affine transformation is presented such as rotation in depth of a face. The signature at the C2 level is not invariant to rotation in depth because the translation invariant global templates at the level of C2 are not invariant to a global non affine transformation. However, an additional layer (S3 and C3) can store a set of class-specific template transformations and provide the required class-specific approximate invariance (see Figures 8 and 9). Notice that invariance to translations and other affine – and generic – transformation up to that level is a significant advantage (observation by J. Leibo).

4. Spectral Properties of Optimal Invariant Templates (with J. Mutch)

In this chapter I consider the question of whether invariances represented in the templatebooks may determine the tuning of neurons by assuming appropriate synaptic plasticity rules. The direct route to answering this question would be to analyze the properties of the learning rule when the inputs are the elements of the templatebooks (at one stage of the hierarchy).

We outline in this section a theory without mathematical details (which can wait) that links optimal invariant templates to tuning of cells in different areas through the spectral properties of the templatebooks. The motivation for analyzing spectral properties is that several plausible Hebbian rules for synaptic plasticity are known to effectively select principal components of the inputs.

Let us start with the following observations[5]:

[5] I could directly consider the elements of the Lie algebra of the continuous group of affine transformations and consider their spectrum. From this point of view translation in x corresponds to $\frac{d}{dx}$. Note that for elements of the algebra which are symmetric matrices A then the spectrum of A and the spectrum of $T = e^{At}$ coincide. This is not true if A is not symmetric and thus is not true for translation

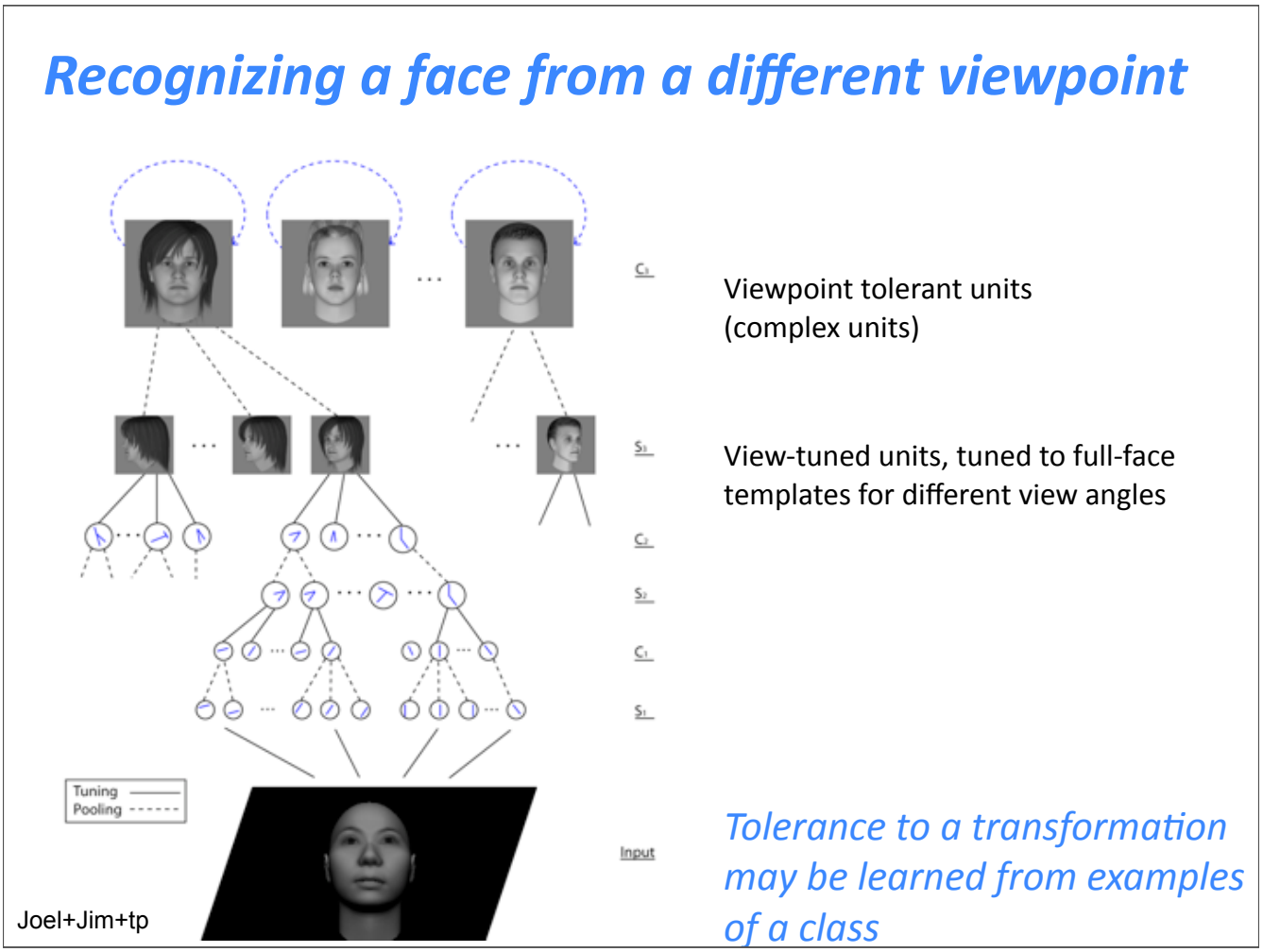


FIGURE 8. *The system with a S3 C3 layer approximatively invariant to face-specific rotations in depth. From [13]. Estimation of identity invariant viewpoint could be performed by a linear classifier receiving inputs from the S3 units with weights to be learned in a supervised training phase.*

- the left singular eigenvectors of the templatebook do not depend on the order of the columns in the templatebook[6];
- the retina performs both a DOG-like spatial filtering operation (Laplacian of a Gaussian) as well as a high-pass filtering in time roughly similar to a time derivative.

The key point here is that, without the time derivative, the templatebooks $\mathbb{T}$ "learned" from one or more transformation sequences – literally storing the frames of a movie – only depend on the statistics of the images and not on the transformation. The temporal derivative however performs the *magic* here: the templatebook then depends on the transformation sequence and so does its spectrum. This can be seen easily in the case of translation where the temporal

[6]The covariance matrix $Q = TT^T$ does not change when columns of T are permuted since $Q = TPP^TT = TT^T$.

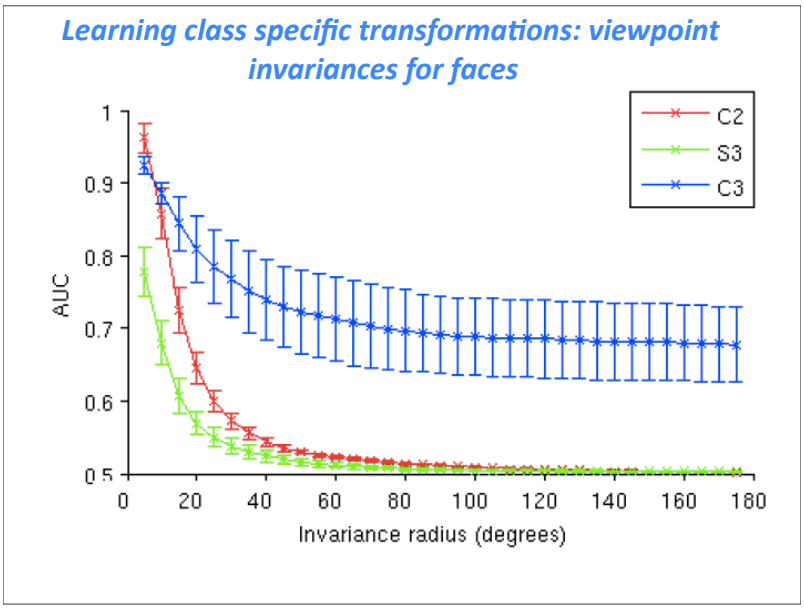


FIGURE 9. *The ordinate shows the AUC obtained for the task of recognizing an individual novel object despite changes in viewpoint. The model was never tested using the same images that were used to produce S3/C3 templates. A simple correlation-based nearest-neighbor classifier must rank all images of 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 naturalpproxi 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.* From [13].

derivative act as a selection rule that prefers orientations orthogonal to the direction of motion. Consider the effect of the time derivative over the movie generated by the translation of an image $f(x - vt)$, where x, v are vectors in $\mathbb{R}^2$:

$$\frac{\mathrm{d}\, f}{\mathrm{d}\, t} = \nabla f \cdot v. \tag{14}$$

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

$$\frac{\mathrm{d}\, f}{\mathrm{d}\, t} = \frac{\partial f}{\partial x} v_x. \tag{15}$$

Thus the effect of the motion in the x direction suppresses spatial changes in y, eg spatial frequences in ω_y, and enhances components orthogonal to the direction of motion. 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*: $i\omega_x F(\omega_x, \omega_y)$. Notice that spatial

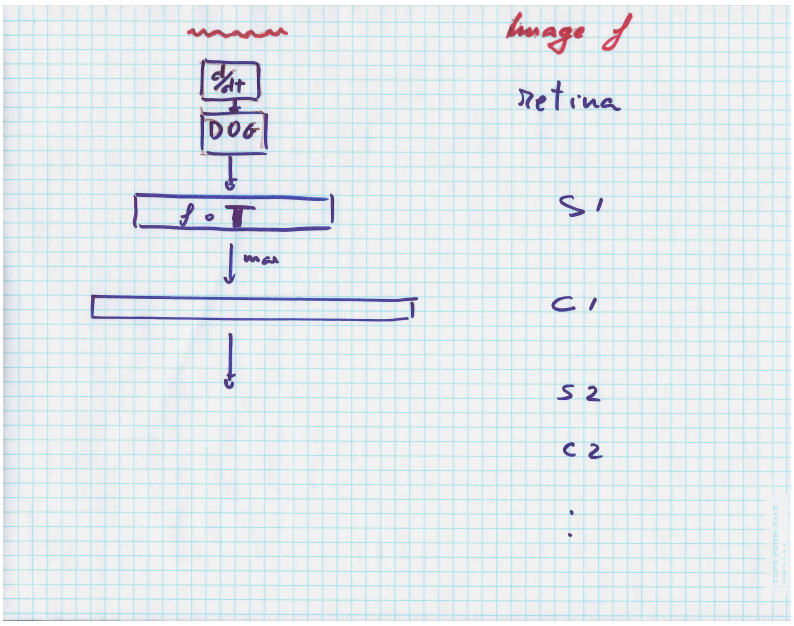


FIGURE 10. *The sequence of processing stage from the retina (with spatial (DOG) and temporal d/dt derivative-like filtering to V1 and V2. More in general, instead of d/dt, I think of a filter such as $(\beta + d/dt)\nabla I(x, y, t)$ which correspond to the power spectrum in the Fourier domain given by $(\beta^2 + \omega_t^2)F(\omega_x, \omega_y, \omega_t)$ with F being the Fourier transform of $\nabla I(x, y, t)$.*

and temporal filtering commute in this linear framework, so their order (in the retina) is not important for the analysis. *The above argument is true not only for translations but for other motions on the plane.* From now on, we assume the pipeline of Figure 10.

Consider now the Lie algebra associated with the affine group on the plane $Aff(2, \mathbb{R}) = GL(2, \mathbb{R}) \times \mathbb{R}^2$. The Lie algebra corresponds to the first order differential equation

$$\frac{\mathrm{d}\, x}{\mathrm{d}\, t} = Ax(t) + x_0 \tag{16}$$

where x, x_0 are vectors in $\mathbb{R}^2$, A is the 2×2 matrix which is here the generator of the Lie group of affine transformations and t is the parameter controlling rotations and scalings. it can be interpreted as time of the transformation. Finite transformations, that is elements of $Aff(2, \mathbb{R})$, are solutions of Equation 16:

$$x(t) = e^{At}[x(0)] + x_0 \int_0^t e^{At} dt]. \tag{17}$$

Because of our assumptions about learning invariances (see section 2), 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 in terms of x_0, A. For $x_0 \neq 0,\ A = 0$ the solution is just a straight line in the direction of x_0. When $x_0 = 0,\ A \neq 0$ the origin is a singular point of the flow. For other initial conditions, the eigenvalues of A determine the structure of the phase space. In particular, we can associate Lie derivatives to the various linear

transformations (for orthographic projection) which transform conic sections (ellipse, parabola, hyperbola, pair of straight lines etc.) into conic sections, leaving the type of conic invariant (thus a circle viewed obliquely will be seen as an ellipse but can still be recognized as a circle).

For instance the Lie derivatives $\mathcal{L}_x = \frac{\partial}{\partial x}$ and $\mathcal{L}_y = \frac{\partial}{\partial y}$ represents translations in x and y respectively; the associated trajectories are families of horizontal and vertical straight lines; the associated features are orthogonal edges (see earlier). In a similar way the Lie derivative associated with positive rotations is $\mathcal{L}_r = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$ represents translations in x and y respectively; the associated trajectories are families of concentric circles; the associated features are a star of radial lines. The Lie derivative associated with the dilatation group is $\mathcal{L}_s = x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$; the associated trajectories are families of radial lines; the associated features are concentric circles. Spiral and hyperbola trajectories are induced by still other Lie derivatives. In general, affine transformations correspond to linear combinations of the infinitesimal generators of the general linear group ($\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, x\frac{\partial}{\partial x}, y\frac{\partial}{\partial x}, x\frac{\partial}{\partial y}, y\frac{\partial}{\partial y}$). We have the following *selection rule*:

Lemma 4. Selection Rule *Assume that a templatebook is obtained after the $\nabla^2 G \circ \frac{\partial}{\partial t}$ filtering of a "video" generated by a transformation which is a subgroup of the affine group $Aff(2, \mathbb{R})$. Then the components in the image spectrum orthogonal to the trajectories of the transformations are preferentially enhanced.*

Remarks

- It may be possible to look at the dynamical system comprising the affine transformations described in their infinitesimal form (see above) and the dynamic of learning (see next section) and infer qualitative properties and in particular invariance directly from such and analysis. Thus it is an interesting *open problem* whether one could develop a direct analysis from transformations to receptive fields using tools from dynamical systems [7].
- For reference I collect in the Appendix (section 7.5) a few notes about transformations and spectral properties of them.
- The hypothesis explored here – PCA of the data matrix – corresponds to maximize the norm of the time derivative of the input patterns (or more precisely a high-pass filtered version of it), because of the temporal properties of the retina. This is related to – but almost the opposite of – the "slowness" principle proposed by Wiskott ([2, 33]) and made precise by Andreas Maurer.

4.1. **Optimizing signatures: the antislowness principle.** A continuous image sequence $I(x, y; t)$ is filtered by the retina in space and time to yield (in Fourier domain) $F(\omega_x, \omega_y; \omega_t) = -(\omega_x^2 +$

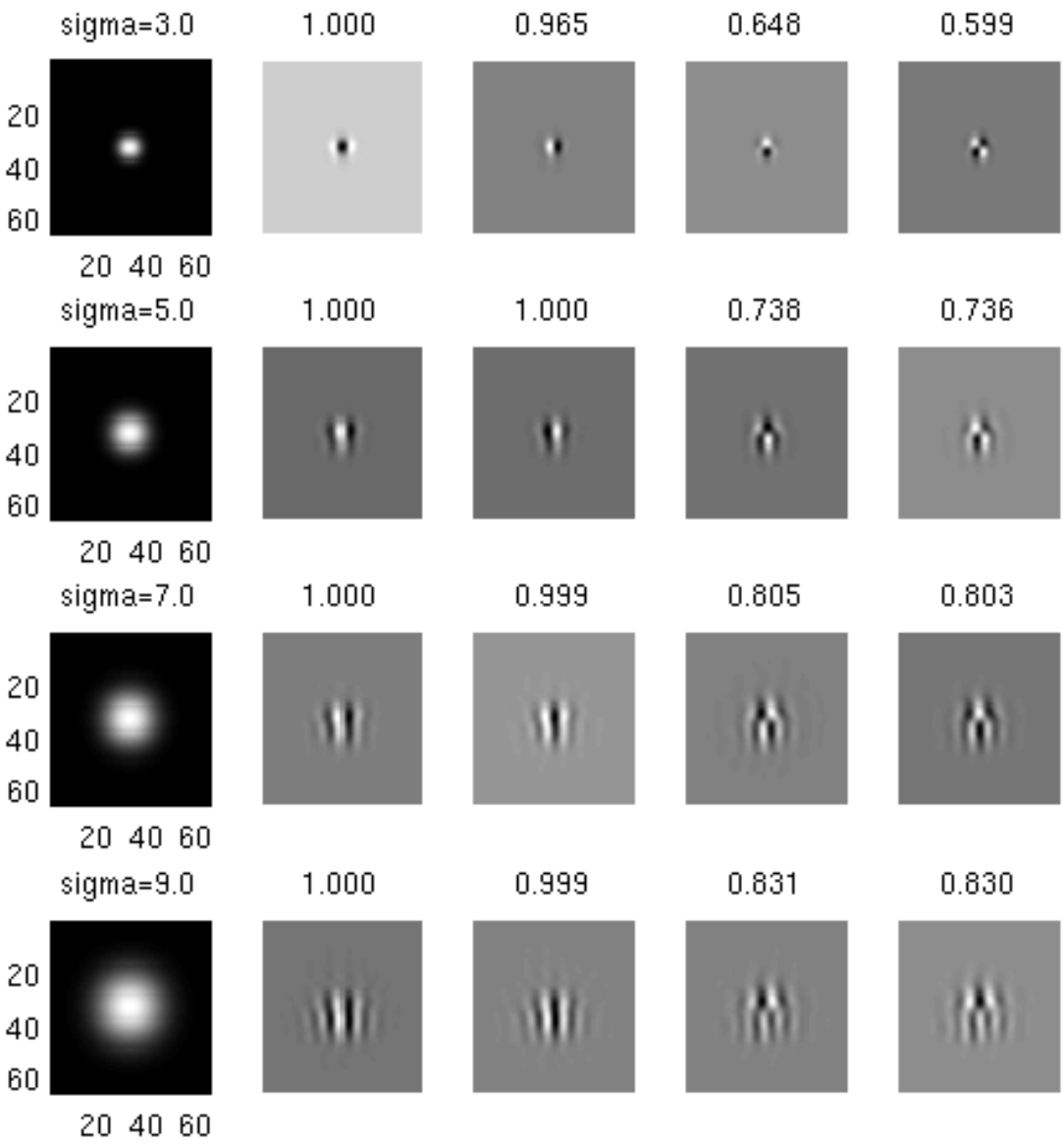


FIGURE 11. *Principal components of the templatebook obtained from (rightwards) translations of natural images looked at through a Gaussian aperture (after DOG filtering and temporal derivative). 10 natural images x 255 translations each. The stratification theorem together with the remarks on spectral properties of transformations implies that Gabor frames should be obtained in the first layer with small apertures. Figures 15 and 18 show some simulations showing that Gabor-like receptive field are generated primarily by transformations rather than by the statistics of natural images.*

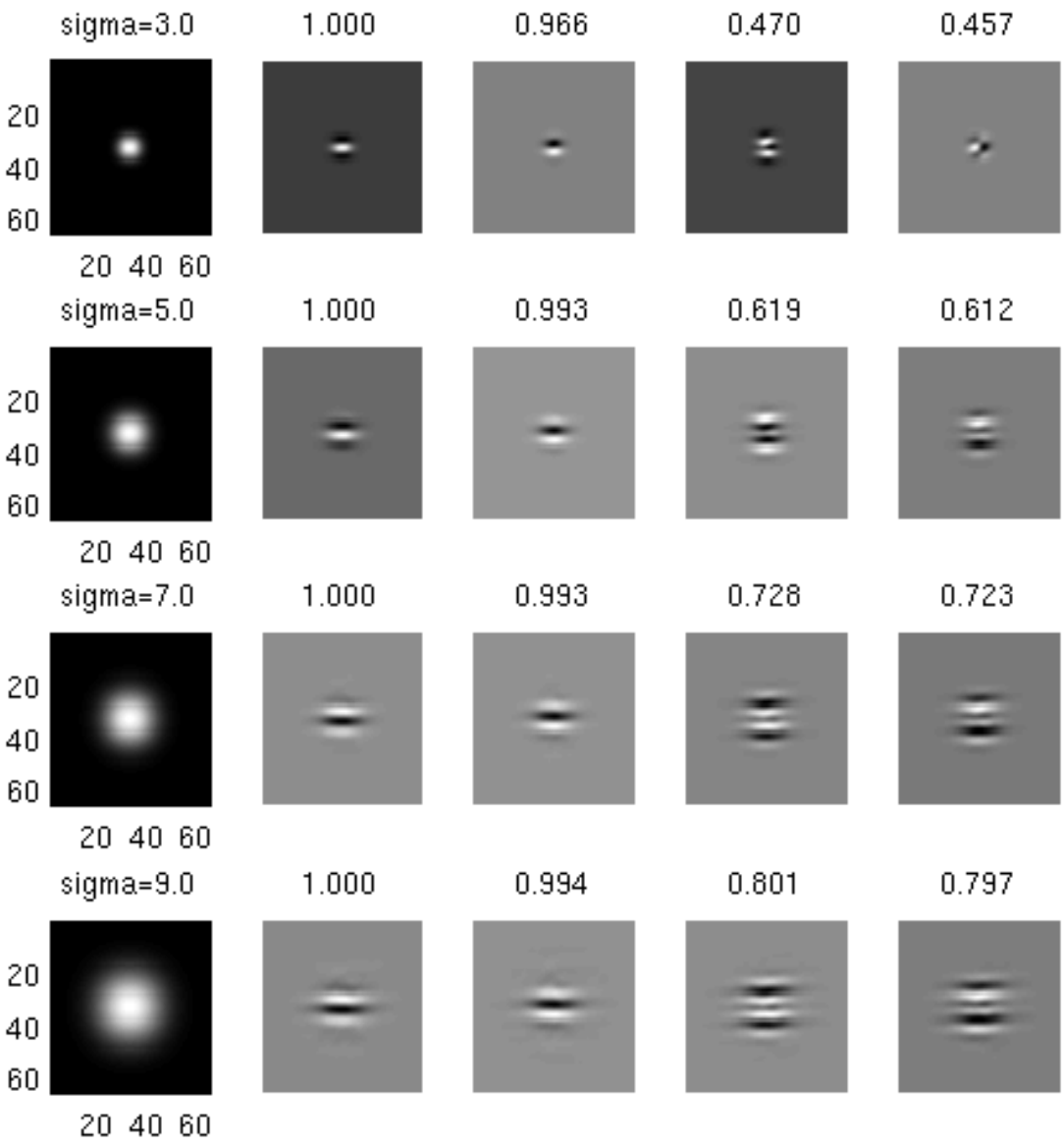


FIGURE 12. *Principal components of the templatebook obtained from vertical translations of natural images looked at through a Gaussian aperture (after DOG filtering and temporal derivative). 10 natural images x 255 translations each.*

$\omega_y^2)G(\omega_x, \omega_y)H(\omega_t)I(\omega_x, \omega_y; \omega_t)$. Assume that

$$H(\omega_t) = (\beta + i\omega_t)\Theta_{\omega^*} \tag{18}$$

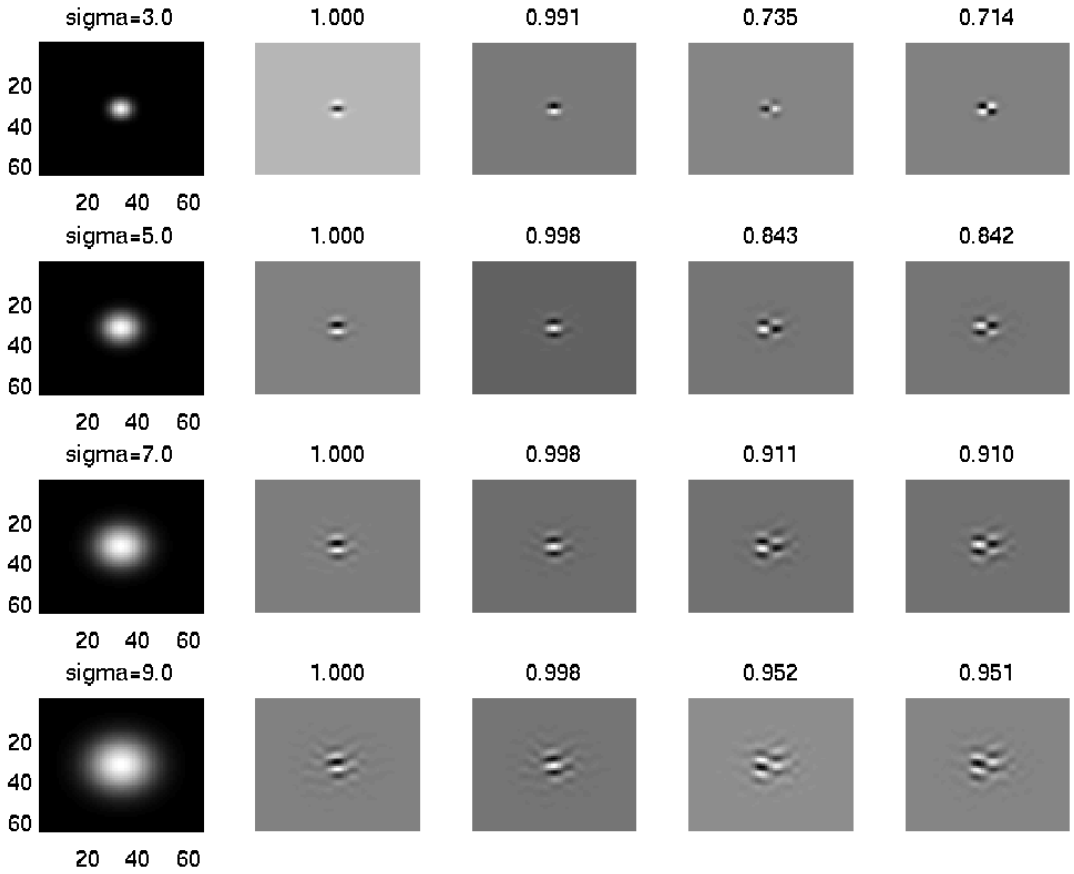


FIGURE 13. *Principal components of the templatebook obtained from vertical translations of random noise images looked at through a Gaussian aperture (after DOG filtering and temporal derivative). 10 natural images x 190 translations each. This suggests that the PCA do not depend much on statistics of natural images.*

where $Theta_{\omega^*}$ is bandpass (eg equal 1 up to $|\omega| = \omega^*$ because the image is low pass, see section 2.1).

Consider now realizations f_i of the process f – representing rows of the templatebook. Taking principal components of the realizations f_i, corresponds to maximize the empirical functional

$$L(P) = \frac{1}{m} \sum_{1}^{m} \beta ||P f_i||^2 \tag{19}$$

wrt P, where P is in the class of d-dimensional orthogonal projections.

We use the following theorem to show that in fact a solution of the maximization problem above can be obtained by projecting onto a dominant eigenspace of L.

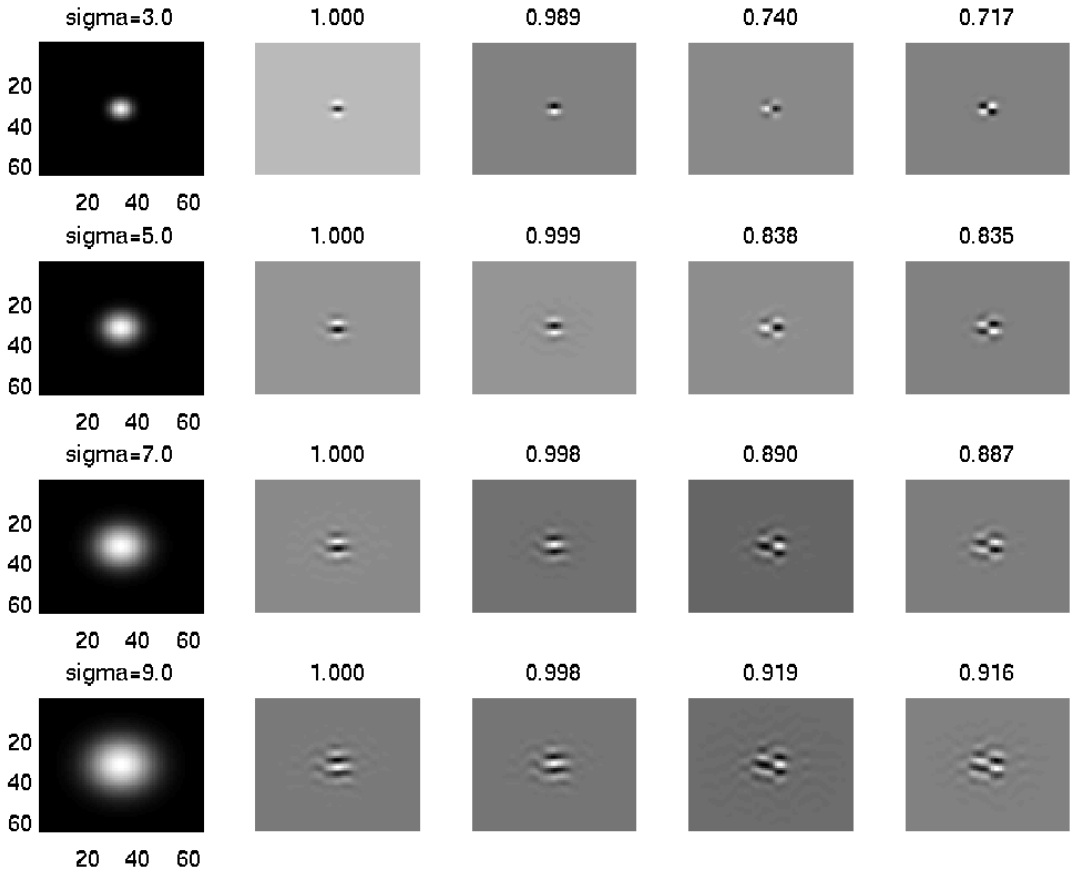


FIGURE 14. *Principal components of the templatebook obtained from vertical translations of RDI (Random Dots Image) looked at through a Gaussian aperture (after DOG filtering and temporal derivative). 10 natural images x 190 translations each. In all these noise simulations, the appearance of SV which look like "end stopped" receptive fields is more reliable.*

Theorem 4. *(Maurer) Suppose that there are d eigenvalues $\lambda_1, \cdots, \lambda_d$ of L so that they are larger than all other eigenvalues; and that e_i is the sequence of associated eigenvectors. Then*

$$max_P L(P) = \sum_i^d \lambda_i \tag{20}$$

the maximum being attained when P is the orthogonal projection onto the span of e_i, $I = 1, \cdots, d$.

As I will discuss later, if learning at the level of the receptive fields of the set of the simple cells which are pooled by one complex cell follows Oja's rule, then their receptive fields cells reflect the Principal Components of the associated templatebook.

Notice that $L(P)$ corresponds to

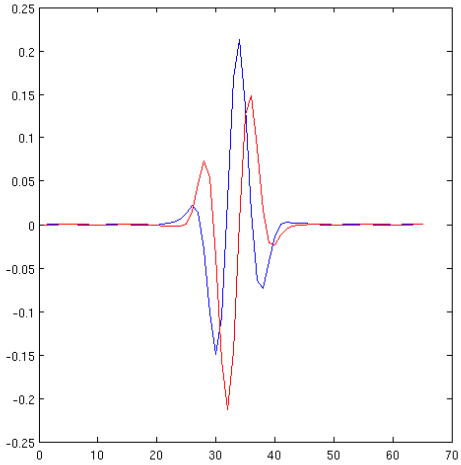


FIGURE 15. *A vertical slice through the 1st and 2nd singular values of aperture sigma=5 in Figure 12.*

(21) $$\hat{L}(P) = \beta C_T + C_{\dot{T}}$$

where C_T and $C_{\dot{T}}$ are the covariance matrices of the processes t and $\dot{t}$ respectively. Notice that the two term sum unlike the slowness case of Maurer. Finding the PCA of the retinallly filtered image in fact projects over projections that maximize changes. In our framework, this is exactly right: invariance follows from pooling together in the same complex cell projections with large changes.

Remarks

- Notice that changing $H(\omega_t) = (\beta + i\omega_t)\Theta_{\omega^*}$ into $\hat{H}(\omega_t) = (\beta - i\omega_t)\Theta_{\omega^*}$ does not change the analysis.
- Remember that the translation subgroup as well as the rotation scaling subgroups are each abelian subgroups[7]. Fourier analysis can be used then to unify the concept of multiresolution analysis in each of the cases. The elementary functions of abelian groups are the *characters*. For our purposes a character χ is a continuous complex-valued function $\chi : G \to \mathbb{C}$ with

(22) $$\chi(x + y) = \chi(x) \cdot \chi(y), \quad |\chi(x)| = 1.$$

In this language, arbitrary functions may be expanded into a superposition of characters which is a Fourier expansion for abelian groups. The Fourier transform over locally compact Abelian groups is an *isometry* and hence it conserves energy (we use this property with the *energy* aggregation function). This means also that the same analysis and

[7]Each of the translation and rotation subgroups forms a representation of the (abelian) additive group since $T(a)T(b) = T(a + b) = T(b)T(a)$ and $R(a)R(b) = R(a + b) = R(b)R(a)$, whereas for dilations the group is multiplicative (F. Anselmi).

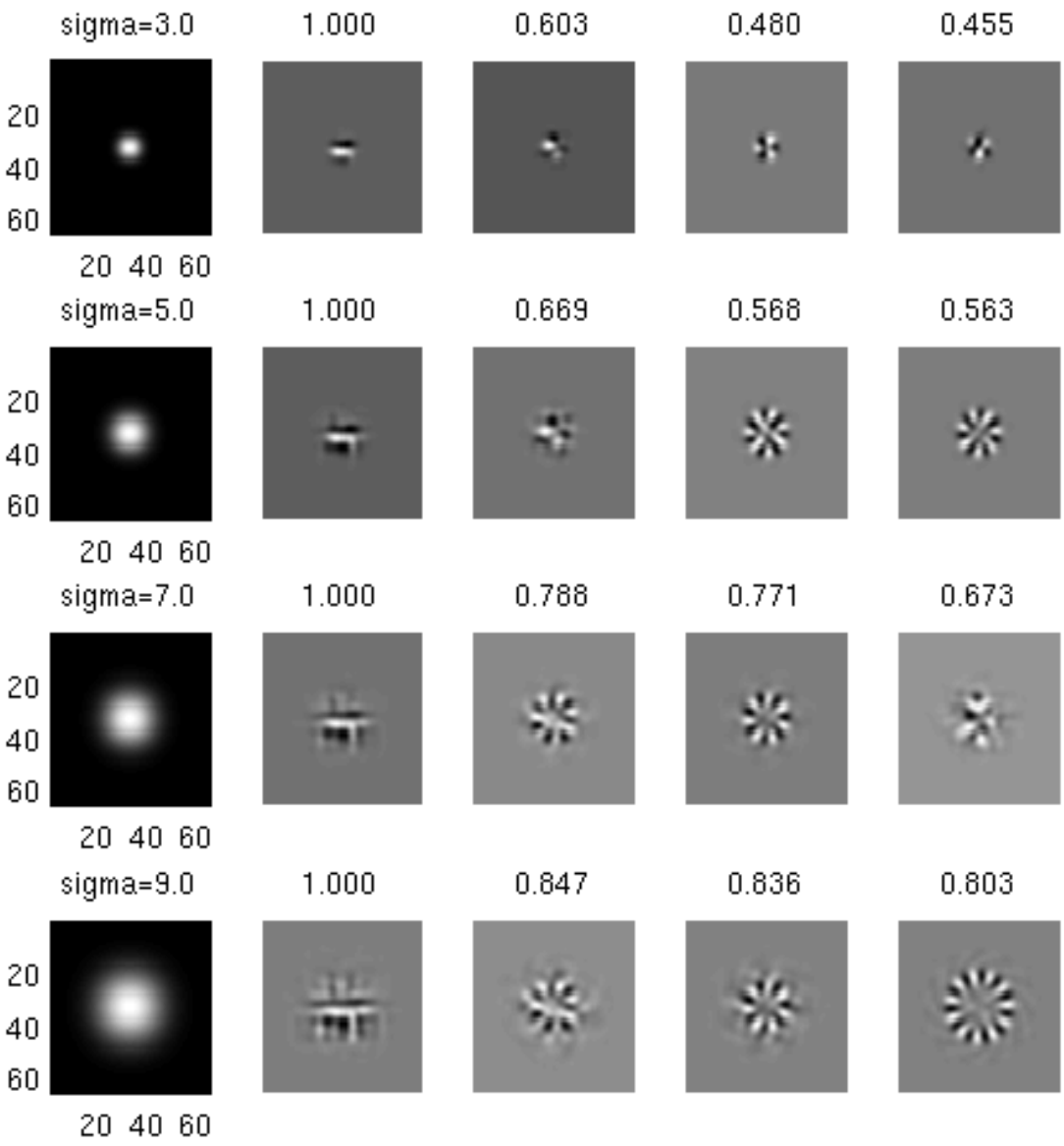


FIGURE 16. *Principal components for a rotating natural image. The rotation is around the center of the aperture.*

the same spectral properties found for translation extend to scaling and rotations and combinations thereof.

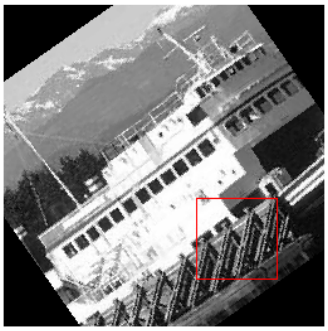


FIGURE 17. *Aperture "looking" at a natural image rotating off-center.*

4.2. **PCAs, Gabor frames and Gabor wavelets.** Let us consider the first layer (corresponding to V1 in the theory) and dealing with translations. The previous subsection together with Appendix 7.5 shows that the top (complex-valued) eigenfunction associated with translation seen through a Gaussian window is a Gabor frame

$$\phi_{\theta,\xi,\eta}(x,y) = e^{-\frac{(x_\theta-\xi_\theta)^2}{2\sigma_x^2}} e^{-\frac{(y_\theta-\eta_\theta)^2}{2\sigma_y^2}} e^{i2\pi x_\theta \omega}$$

where the basic frequency ω depends on the scale parameter of the associated retinal DOG filtering and the σ are fixed. Eigenvectors with lower eigenvalues will have multiple of the basic frequency.

Let us now *assume* based on biological plausibility that

(1) at each position there are several retinal ganglion cells each associated with DOG filtering of different size
(2) the size of the Gaussian aperture of a simple- complex cell is proportional to the size of the DOG filters associated with the ganglion cells-LGN inputs pooled by the simple-complex cell (this corresponds to assuming that complex cells pool the same number of afferent inputs from the LGN, independently of DOG size).

The consequence of these assumptions is that σ becomes inversely dependent on ω, eg at each position the top eigenfunction for each size gives a system of Gabor wavelets

$$\phi_{\theta,\xi,\eta}(x,y) = e^{-\frac{(x_\theta-\xi_\theta)^2}{2\frac{c_x}{\omega^2}}} e^{-\frac{(y_\theta-\eta_\theta)^2}{2\frac{c_y}{\omega^2}}} e^{i2\pi x_\theta \omega}. \tag{23}$$

5. TOWARDS A THEORY: PUTTING EVERYTHING TOGETHER

In this section I focus on the module of Figure 23 which is repeated across and through layers in architectures such as the architecture of Figure 2. The module is of course a cartoon and its details should not be taken too seriously.

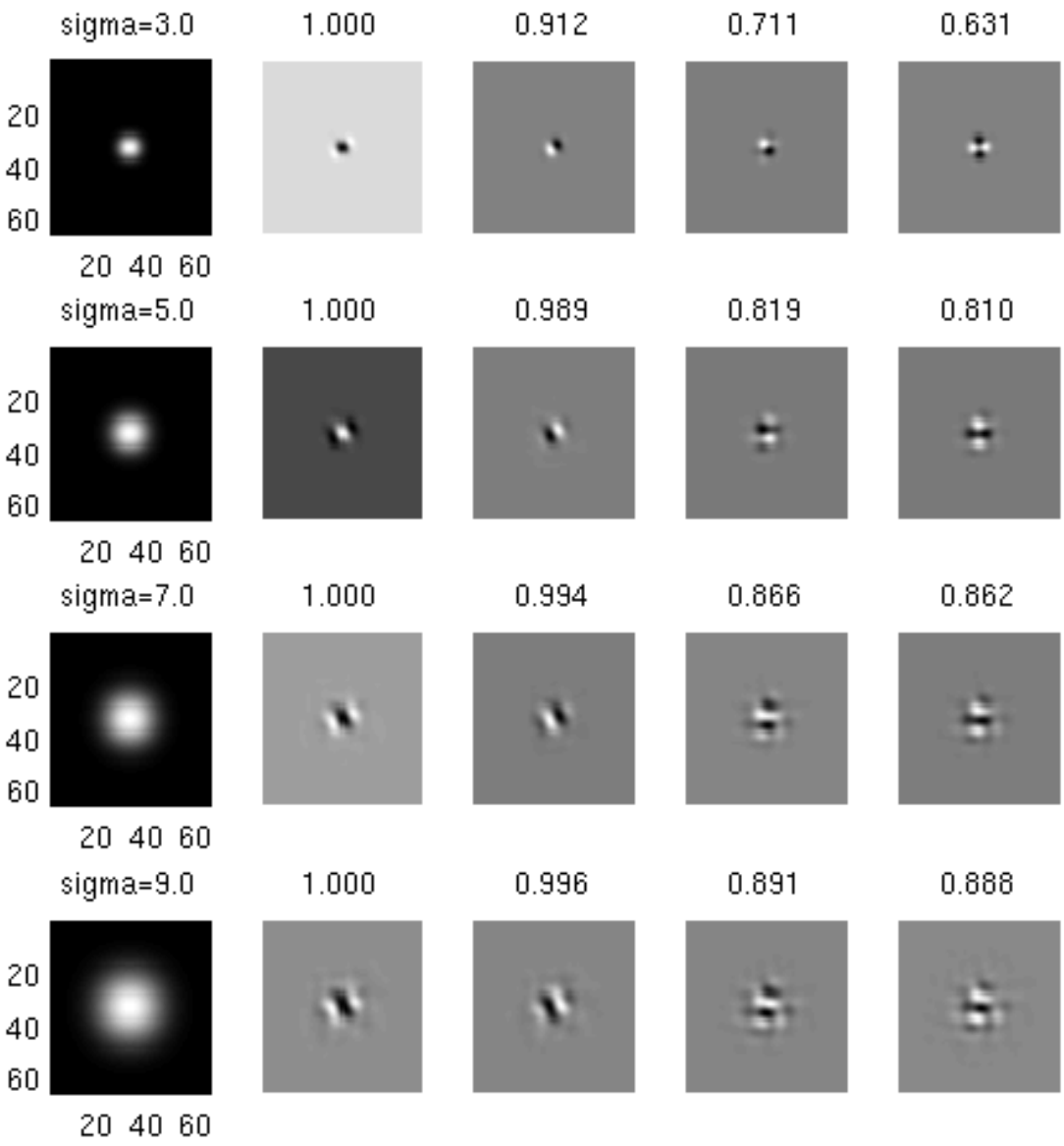


FIGURE 18. *Principal components of the templatebook obtained from rotating natural images such as that of Figure 16 when they are viewed through a Gaussian aperture which does not include the center of rotation (after DOG filtering and temporal derivative).*

5.0.1. *Learning rule and receptive fields (with J. Mutch).* The algorithm outlined earlier in which transformations are "learned" by memorizing sequences of a patch undergoing a transformation is a complete algorithm similar to the existing HMAX (in which S2 tunings are learned by sampling and memorizing random patches of images). A biologically more plausible online

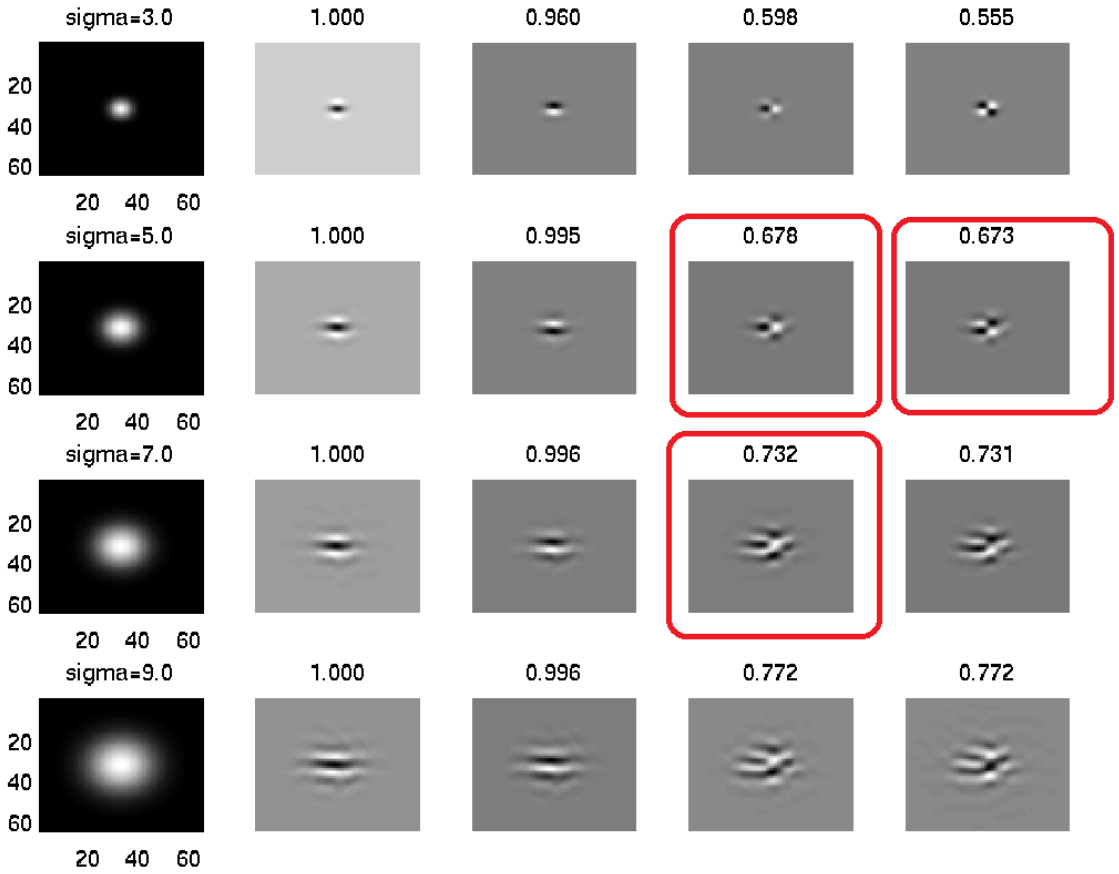


FIGURE 19. *Some of principal components of the templatebook look like the end-stopped cells of Hubel and Wiesel.*

learning rule would however be 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 Hebb-like rules. Notice that Hebb-like rules may lead synapses at the level of simple-complex cells to match their tuning to the eigenvectors of the templatebooks (Hebb-like rules turn out 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.

In particular, let us consider Oja's rule [10] 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.

Oja's rule defines the change in presynaptic weights $\mathbf{w}$ given the output response $\mathbf{y}$ of a neuron to its inputs to be

$$\Delta \mathbf{w} \;=\; \mathbf{w}_{n+1} - \mathbf{w}_n \;=\; \eta\, y_n(\mathbf{x}_n - y_n \mathbf{w}_n) \tag{24}$$

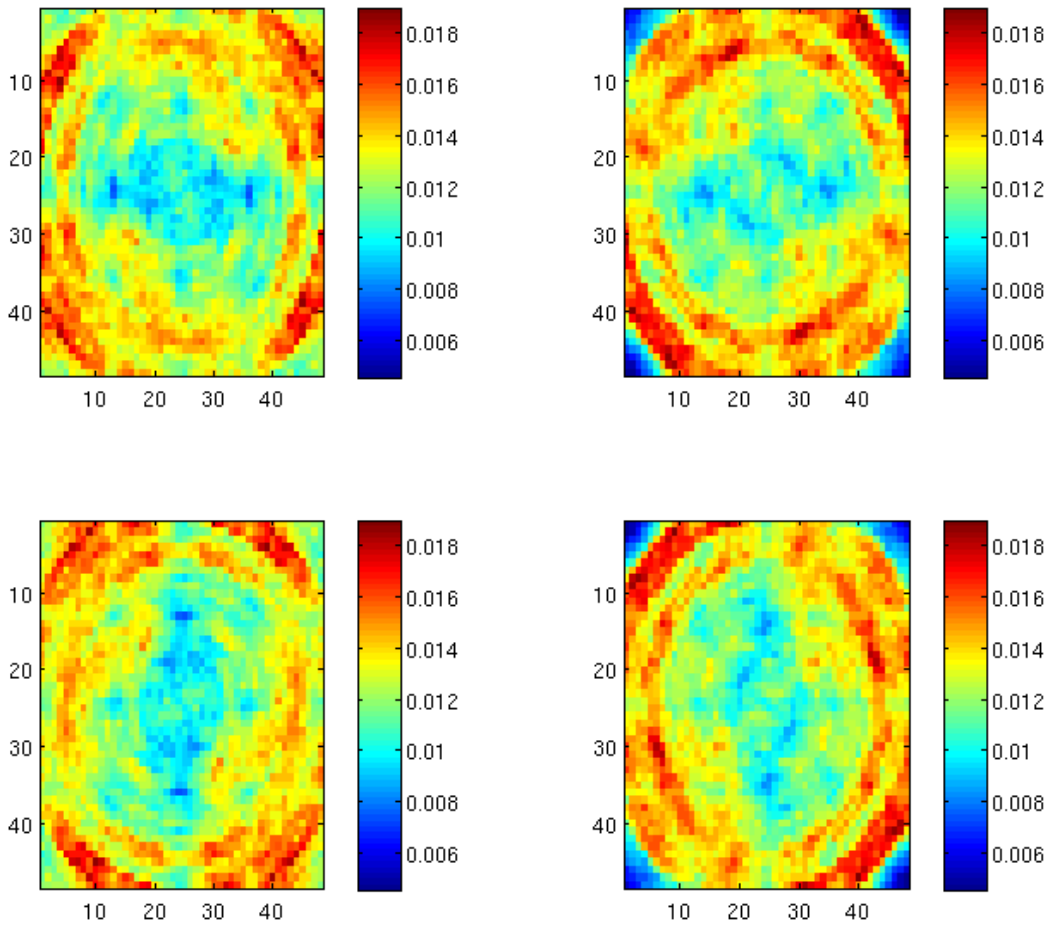


FIGURE 20. *Average of the absolute value of the *change* in C1 units for rotations over 360 degrees.*

where is the "learning rate" and $y = \mathbf{w}^{\mathbf{T}}\mathbf{x}$. Notice that the equation follows from expanding to the first order Hebb rule normalized to avoid divergence of the weights. Hebb's original rule, which states in conceptual terms that "neurons that fire together, wire together", is written as $\Delta\mathbf{w} = \eta\, y(\mathbf{x}_n)\mathbf{x}_n$. Hebb's rule has synaptic weights approaching 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:

$$w_i(n+1) = \frac{w_i + \eta\, y(\mathbf{x}) x_i}{\left(\sum_{j=1}^{m} [w_j + \eta\, y(\mathbf{x}) x_j]^p\right)^{1/p}} \tag{25}$$

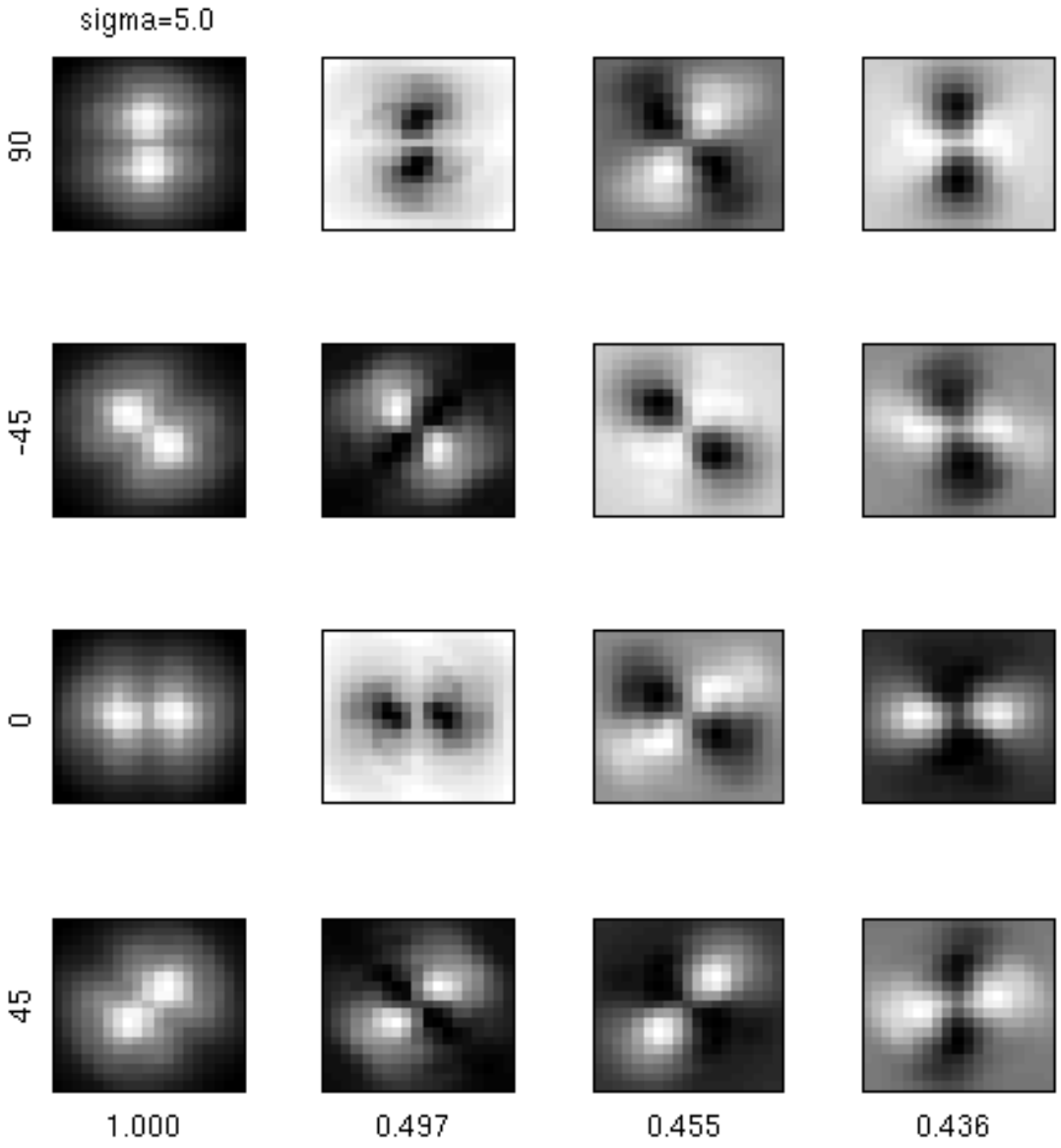


FIGURE 21. *Rotation centered in the RF. 10 images x 360 steps. C1 units of 4 orientations. DOG + D/DT + S1 + C1 + aperture + SVD. 1st four singular values for each aperture.*

of which Oja's rule is an approximation.

Notice that several theoretical papers on Hebbian learning rules, showed 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

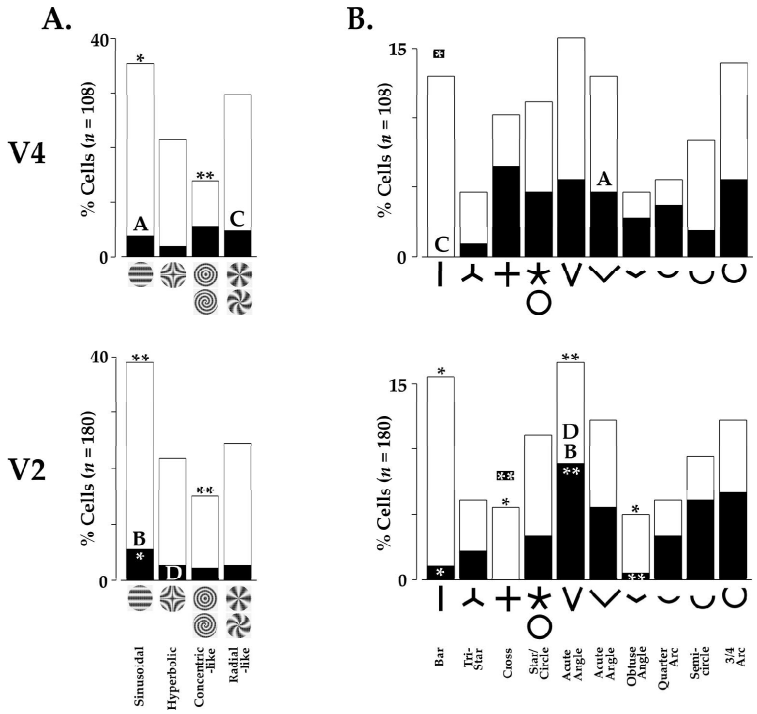


FIGURE 22. *The effectiveness of the various stimulus subclasses for V4 vs. V2. Each cell from either area was classified according the subclass to which its most effective grating or contour stimulus belonged. The resulting distributions are shown here for grating stimuli (panel A) or contour stimuli (panel B) for both V4 (top row) and V2. See [5].*

control of synaptic plasticity – corresponding to the normalizing term in Oja equation – ([31]) is experimentally well established.

The rules above find the PCA with the largest eigenvalue (see Appendix 7.5.9). I conjecture that a variation of Oja's flow above, with appropriate circuitry to support it, 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 the following theorem:

Theorem 5. *The Oja flow (Equation 24) generates synaptic weights that converge to the two top real eigenvectors of the input patterns covariance matrix, that is the covariance matrix of the templatebook.*

The theorem does not by itself imply the details of what we need. Consider for instance a max aggregation function. We would need transformed (for instance, translated) versions of the nth principal component to become the tuning of the simple cells pooled by one complex cell. A possible scenario is that a Hebb-like rule determines during development tunings of simple cells following, say, the first principal component with different phase shifts. Then Foldiak-type learning would wire the "correct" simple cells to one complex cell.

There is however a key observation (by J. Mutch) that allows to link this result with a more elegant and more plausible scheme for developmental learning of simple-complex cells. The observation is that the top two PCA for the translation case are a quadrature pair and that this should be true for the other subgroups of the affine group since the characters are always

Fourier components. It follows that the *energy* aggregation function is automatically invariant (because $|e^{i\omega nx+\theta}| = 1$) to the transformation. Thus the conjecture is that online learning from 'objects transforming will induce tunings of "simple cells" corresponding to the quadrature pair (see Figure 15). Following section 4.2 and Equation 23 the conjecture then is

Theorem 6. *Linking conjecture (Mutch and Poggio): If learning at the level of the synapses between LGN inputs and "simple cell" dendritic branches pooled by one complex cell follows an Oja-like rule, then their receptive fields cells tuning will 1) reflect the Principal Components of the associated templatebook 2) the top two real-valued PCA (with the same largest eigenvalue) for each aperture are Gabor frames in quadrature pair, 3) at each position over a set of apertures with different sizes the same PCA form a set of Gabor wavelets and 4) their wiring implements invariance via an energy aggregation function that satisfies the invariance lemma.*

Notice that small changes in the Oja equations give an online algorithm for computing ICAs instead of PCAs. What is best theoretically, associated properties and what is true biologically are all open questions$^{\oplus}$. It may well be that the same learning rule determines the pooling and the tuning of the simple cells receptive fields. From this point of view it seems possible (but not necessary) that a simple cell may be just be a group of inputs on a dendritic branch of a complex cell. Thus a version of the architecture, with this learning rule, may link the spectral properties of $\mathbb{T}$ to the tuning of the simple units. Figure 23 shows a cartoon of the of the model.

Let us summarize the main implications of this section in terms of templates, signatures and simple+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 15 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 of a complex cell to stimuli sequences corresponding to the rows of the $\mathbb{T}$ should induce, through the learning rule, simple cell tunings corresponding to the two PCA in quadrature pair of Figure 15. Simple cells tuned to this Principal Component (it one component in the complex domain) would be pooled by the same complex cell. I think that the learning rule may be complemented by interactions between complex cells pooling different principal components to achieve the development of a set of complex cells capable of providing a discriminative and invariant local signature with the correct lattice density in space and scale.

As noted by J. Mutch, the subgroup of translations is a 2parameter group (translations in x, y); the subgroup of rotations and dilations is also a two parameters group (ρ, θ).

6. Discussion

In this section I will first summarize the main ideas emerging of the theory, compare the new theory with the old model, list potential problems and weaknesses (in "What is under the carpet") and finally discuss directions for future research.

6.1. Summary of the main ideas.

There are several key ideas in the theoretical framework of the paper. There are hypotheses and there are theorems.

(1) First, I conjecture that the sample complexity of object recognition is mostly due to geometric image transformations (different viewpoints) and that a main goal of the ventral stream – V1, V2, V4 and IT – is to learn-and-discount image transformations. The most

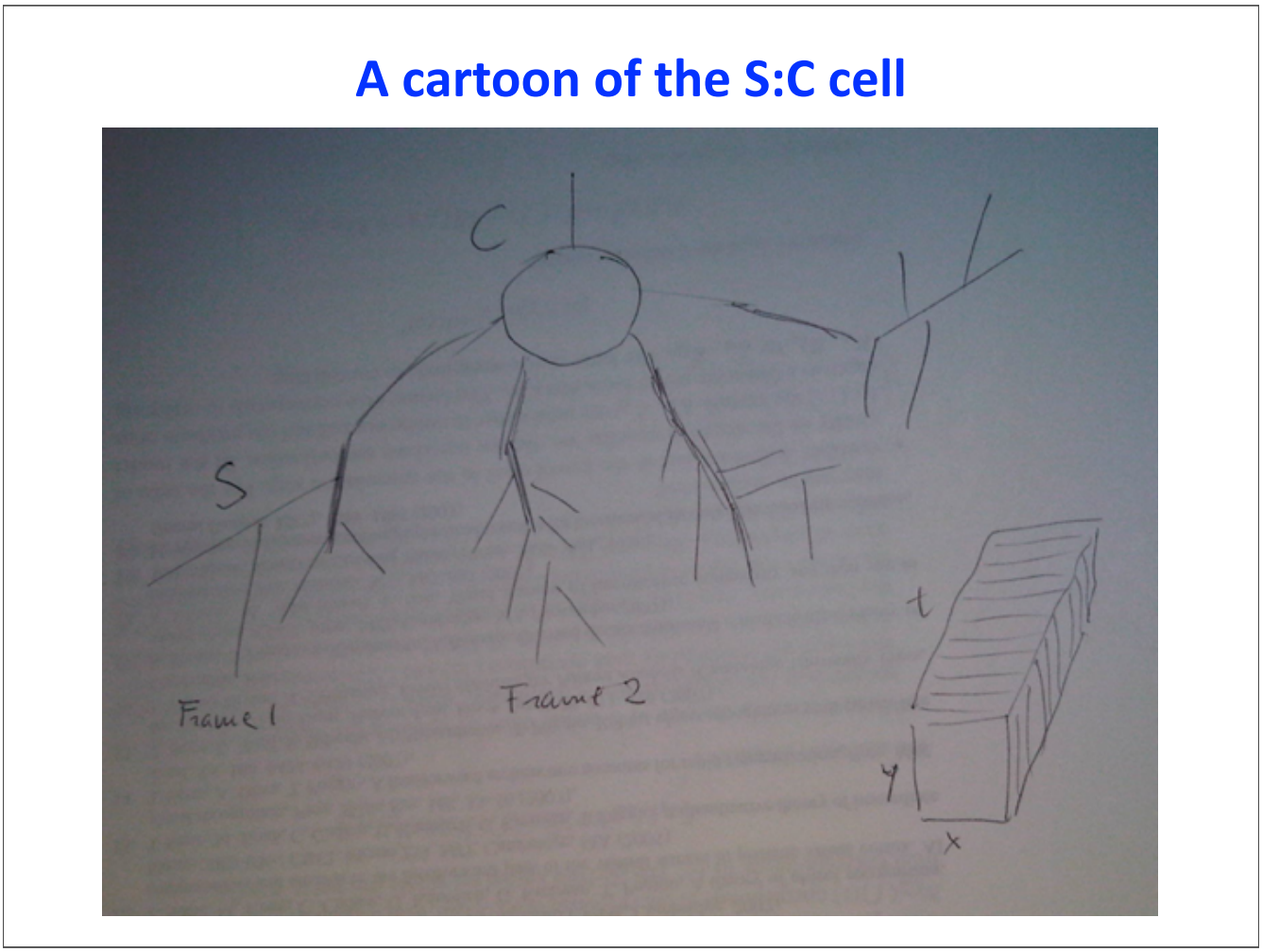


FIGURE 23. *Cartoon of an SC cell with dendrites representing simple cells and the cell body performing complex-like pooling. Notice that the theory suggests that the spatial extent of the receptive field is in general the same for simple cells and for the complex cell that "pools" them. Of course, learning (by a Hebb-like rule) may induce zero weights in some parts of the receptive field. Here the simple cells are represented as dendritic branches of a complex cell. The theory leaves open the question of whether simple cells may instead be independent neurons.*

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 to determine through evolution and 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.

(2) Second, aggregation functions such as the max (in HMAX) ensure that signatures of images are invariant to affine transformations of the image and that this property is preserved from layer to layer.
(3) Third, I assume that there is a hierarchical organization of areas of the ventral stream with increasingly larger receptive fields. The stratification conjecture claims that small apertures determine a stratification of the invariances from translations to full affine in highest layers.
(4) The fourth 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.
(5) The final idea is that spectral properties determine receptive field tuning via Hebbian-like online learning algorithms that converge to the principal components of the inputs.

The theory part of this paper start 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, I 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 (the *affine lemma*), 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 likely to be a good approximation of reality. I then show that memory-based hierarchical architectures are much better at learning transformations than nonhierarchical 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 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*).

In section 4 I speculate on how the properties of the specific areas may be determined by visual experience and continuous plasticity. I characterize 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; raising kittens in a world made of random

dots should yield normal receptive properties (because all types of motion and transformations will be preserved) though a world made of vertical stripes only should affect receptive fields properties;

- the type of transformations that are learned from visual experience depend on the size of the receptive fields (measured in terms of spatial wavelength) and thus on the area (layer in the models) – assuming that the size increases with layers;
- 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;
- the mix of transformations 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 (eg face tuned cells);
- invariance to small transformations in early areas (eg translations in V1) may underly stability of visual perception (suggested by Stu Geman);
- simple cells are likely to be the same population as complex cells, arising from different convergence of the Hebbian learning rule. The input to complex "complex" cells are dendritic branches with simple cell properties;
- the output of the ventral stream is a *G-invariant signature*, eg is a vector that can be used as a key for an associative memory (or of a vector-valued classifier); multiple signatures for classification can be extracted from intermediate areas;
- class-specific modules – such as faces, places and possibly body areas – should exist in IT to process images of object classes;
- the mix of transformations learned determine the properties of the receptive fields – oriented bars in V1+V2, radial and spiral patterns in V4 up to class-specific tuning in AIT (eg 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[8].
- invariance to affine and other transformations can provide the equivalent of generalizing from a single example to "conceptual" invariances;
- the *transfer of invariance* accomplished by the machinery of the templatebooks may be used to implement high level abstractions;
- the effect of small motions of the image should decrease going from V1 to IT – apart from the on-off effects due to temporal derivative-like filtering in the retina and cortex.
- In the preceeding sections I stressed that the statistic of natural images play a secondary 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

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

over the presentations affects the spectral content, eg the correlation matrix and thus the stationary solutions of Hebbian learning.
- In summary, from the assumption of a hierarchy of areas with receptive fields of increasing size the theory predicts that the size of the receptive fields determines which transformations are learned during development and then factored out during normal processing; that the transformation represented in an area determines the tuning of the neurons in the area; and that class-specific transformations are learned and represented at the top of the hierarchy.

6.2. **Extended model and previous model.** So far in this paper, existing hierarchical models of visual cortex – eg HMAX – are reinterpreted and extended in terms of computational architectures which evolved to discount image transformations learned from experience. From this new perspective, I argue that a main goal of cortex is to learn equivalence classes consisting of patches of images (that we call templates), associated together since they are observed in close temporal contiguity – in fact as a temporal sequence – and are therefore likely to represent physical transformations of the same object (or part of the same object). I also conjecture that the hierarchy – and the number of layers in it - is then determined by the need to learn a group of transformations – such as the affine group. I prove that a simple memory-based architecture can learn invariances from the visual environment and can provide invariant codes to higher memory areas. I also discuss the possibility that the size of the receptive fields determines the type of transformations which are learned by different areas of cortex from the natural visual world – from local translations to local rotations and image-plane affine transformations up to almost global translations and viewpoint/pose/expression transformations. Earlier layers would mostly represent local generic transformations such as translation and scale and other similitude transformations. Similar considerations imply that the highest layers may represent class-specific transformations such as rotations in depth of faces or changes in pose of bodies.

- The present Hmax model has been hardwired to deal with 2 generic transformations: translation and scale. Figure 24 shows that the model performance on "pure" translation tasks is perfect, while it declines quickly with viewpoint changes.
- Signatures from several layers can be used by the classifier, possibly under attentional or top-down control, possibly via cortical-pulvinar-cortical connections.
- What matters for recognition is not strong response of a population of neurons (representing a signature) but invariance of the response in order to provide an input invariant as much as possible to the classifier.

6.3. **Invariance to X and estimation of X.** So far I have discussed the problem of recognition as estimating identity or category invariant 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 [9] and model simulations [26]. How is this possible in the framework of the theory of invariant signatures? Consider a specific example. Suppose that the top layer, before the final classifier, has a templatebook recorded under viewpoint transformation of a face. A max operation on the dot products of row i of the templatebook – eg $max_j \quad t_{i,j}$ – provides a number that is (almost) invariant to viewpoint (for new

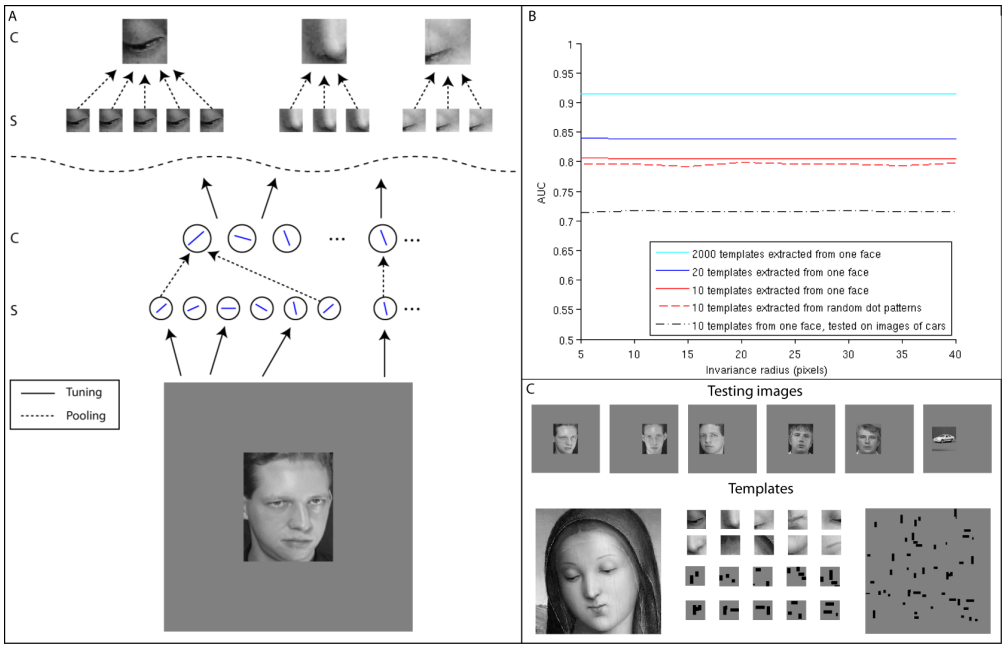


FIGURE 24. *(a) Illustration of a generic hierarchical model of object recognition in the spirit of Hubel and Wiesel. In the first layer (S), the simple units are tuned to oriented edges. Each complex unit in the second (C) layer, pools the first layer units with the same preferred orientation but from different locations in the visual field. In the penultimate layer, cells are tuned to patches of natural images. Each high level C unit pools S cells tuned to the same template replicated at different locations. The image signature computed by the top level C units is then fed into a classifier. (b) Model accuracy: a summary statistic of the ROC curve for classifying test images as either containing the same person's face or a different person's face. AUC ranges from 0.5- indicating chance performance- to 1- indicating perfect performance. We repeat this classification allowing objects translate different distances (pixels). AUC is shown as a function of the invariance range, over which objects could appear. The receptive field of the top-level C units was 256x256 pixels; the faces were approximately 100 pixels horizontally. 10 test images of the target face under slightly variable pose and lighting conditions were used; each was replicated at every position in a radius of 40 pixels. The distractors were 390 different faces presented at the same locations. The simulations were done with a linear correlation classifier using only a single training view (with ten to be associated at each position) presented at the center of the receptive field. Just 10 top-level C units were sufficient for good performance. Each C cell pooled from 3000 S cells optimally tuned to the same stimulus at each location. The black trace shows the results from testing on images of cars. In this case there were 5 views of the target car to be associated at each position and 45 distractor cars replicated at each position. (c) Examples of test images and top level templates. See [17].*

face images), see Figure 7. Suppose instead that all the j components are used as a vector input to a linear classifier – thus the $t_{i,j}$ for fixed i and $j - 1, \cdots, N$ are the "centers". The classifier may be trained in a supervised way to classify identity invariant to pose or pose invariant to identity. Notice that here I do not require that pose and identity transformations have a group structure nor I require that the templates are closed under the set of transformations. Notice also that if the elements of the templatebook at that layer are already completely invariant to X, then an estimate of X following the approach outlined above is not possible.

I think that estimates of pose, viewpoint, expression are especially important (as well as invariance to them). 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, possibly involving attention.

Of all the transformations, pose is probably one of the most important from the evolutionary point of view. It is therefore natural to predict from the theory developed so far

- the existence of face patches dedicated to face identification independent of viewpoint, expression, illumination and even age (all transformations that can be learned approximatively). The same or other face patches – using in part different neurons and circuitry but the same inputs – are also capable of *estimating* age, expression, illumination and viewpoint (independent of identity).
- the existence of a body area dedicated to recognizing independent of body pose but more importantly capable of estimating pose to support answers to questions such as : is this facing towards or facing away? is this jumping or kneeling?. Such an area may be quite large and may be one of the reasons underlying the difference reported in fMRI and physiology studies between animate and inanimate objects.

For a few transformations – say pose or translation – invariance vs estimation could be obtained by using the max operation vs the $argmax$ operation – both operations understood as being over the group (assuming that an underlying group structure exists), without the need of assuming supervised learning at the top layer.

6.4. **What is under the carpet.** Here is a list of potential weaknesses, small and large, with some comments:

- "The theory is too nice to be true". One of the main problems of the theory is that it seems much too elegant – in the sense of physics – for biology.
- Backprojections are not taken into account and they are a very obvious feature of the anatomy, which any real theory should explain$^{\oplus}$. As a (lame) excuse let me mention that is plenty of room in a realistic implementation of the present theory for top-down control signals and circuits, managing learning and possibly fetching signatures from different areas and at different locations in a task-dependent way. A 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
- Subcortical projections, such as, for instance, projections to and from the pulvinar not predicted by the theory. The present theory still is (unfortunately) in the "cortical chauvinism" camp. I hope somebody will rescue it$^{\oplus}$.
- Cortex is organized in a series of layers with specific types of cells and corresponding arborizations and connectivities. The theory does not say anything about

6.5. **Intriguing directions for future research.**

6.5.1. *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 classifer.

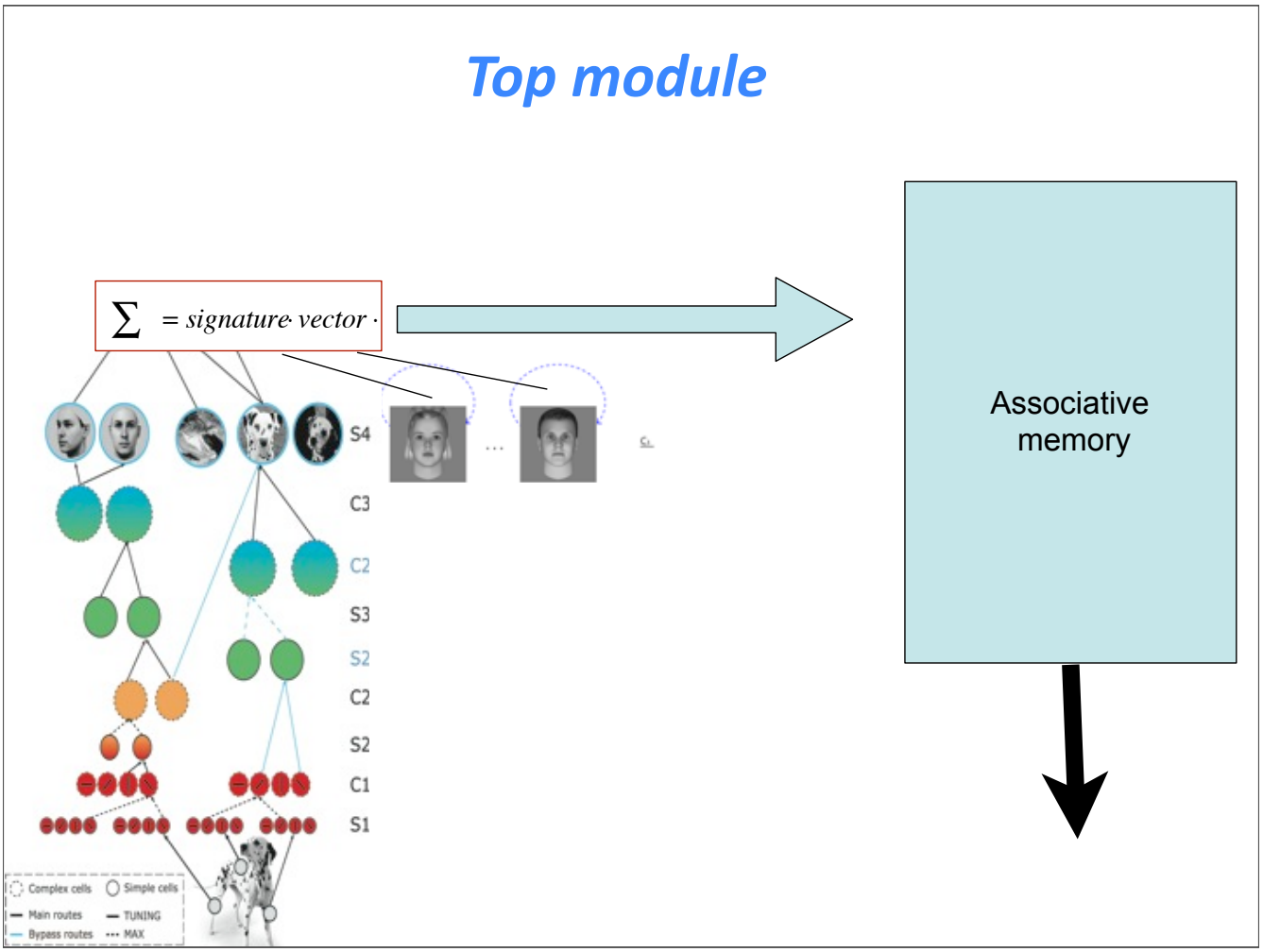


FIGURE 25. *The signature produced at the last level (possibly combined with intermediate level signatures) accesses an associative memory to retrieve information such as a label or an action to be taken. The signature was previously used to store related information, such as a label or similar images or experiences associated with the image.*

Here I would like to extend this view by assuming that the *signature vector* is input to an associative memory so that a number of properties 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 [18, 20]). It would be interesting to connect$^{\oplus}$ these results to estimated capacity of visual memory (Oliva, 2010).

Associative computational modules Consider the architecture studied here as part of an associative computing machine. In addition to the memory-based module described in this paper, one would like to have a few other basic ones. Associations should be able to do binding, such as

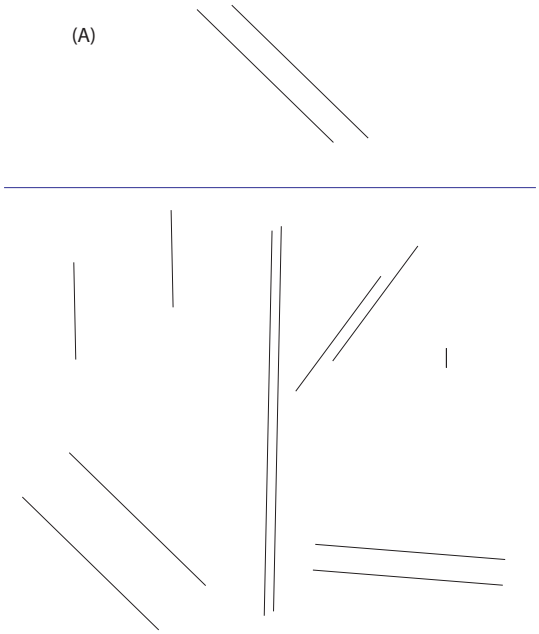


FIGURE 26. For a system which is invariant to affine transformations a single training example (A) allows recognition of all other instances of parallel lines – never seen before.

.

bind A to B. Could we develop a *universal associative machine* which is equivalent to a universal Turing machine and is biologically plausible$^{\oplus}$.

SC stage as association If any of the S cells fires, then the C cell fires. This is an OR operation at run-time and an association during learning. At run time input x retrieves y. The magic of the memory-based module described in section 2 is that y is independent of transformations of x.

Weak labeling by association of video frames Assume to associate together in the top associative module (see Figure 25) images in a video that are contiguous in time (apart when there are clear transitions). This idea (mentioned by Kai Yu) relies on smoothness in time to label via association. It is a very biological semisupervised learning, very much in spirit 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$^{\oplus}$.

6.5.2. *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.
- *Line drawings conjecture* The memory-based module described in this paper should be able to generalize from real images to line drawings when exposed to illumination-dependent transformations of images. This may need to happen at more than one level

in the system, starting with the very first layer (eg V1). Generalizations with respect to recognition of objects invariant to shadows may also be possible.

6.5.3. *Invariance and Perception.* 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 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 and classical (western) music of the transformations of human speech$^{\oplus}$.

6.5.4. *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$^{\oplus}$.

6.5.5. *For philosophers: Is the ventral stream a cortical mirror of the invariances of the physical world? Is the brain mirroring the physics of the 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 much of the computational goals, of the hierarchical organization and of the tuning properties of neurons in visual areas. I have to make an easy joke inspired by the famous line of a close friend of mine: it did not escape my attention that the theory described in this technical report implies that the brain function, structure and properties reflect in a surprising direct way the physics of the visual world.

Acknowledgments I would like to especially thanks Fabio Anselmi, Jim Mutch, Joel Leibo, Lorenzo Rosasco, Steve Smale, Leyla Isik, Owen Lewis, Alan Yuille, Stephane Mallat, Mahadevan for discussions leading to this preprint. Krista Ehinger and Aude Oliva provided to J.L. the images of Figure 1 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, A. Tacchetti, 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] F. Crick, D. Marr, and T. Poggio. An information-processing approach to understanding the visual cortex. In *The Organization of the Cerebral Cortex*, pages 503–533. E.O. Schmitt, F.G. Worden and G.S. Dennis (eds.), MIT Press, Cambridge, MA, 1980.

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

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

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

[5] J. Hedge and D. V. Essen. Selectivity for Complex Shapes in Primate Visual Area V2. *Journal of Neuroscience*, pages 1–6, 2000.

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

[7] M. W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems and Linear Algebra*. Academic Press, 1974.

[8] W. Hoffman. The Lie Algebra of Visual Perception. *Journal of Mathematical Psychology*, (3):65–98, 1966.

[9] 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.

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

[11] J. Koenderink and A. J. van Doorn. Receptive Field Families . *Biological Cybernetics*, (63):291–297, 1990.

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

[13] J. 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.

[14] J. 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.

[15] J. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio. Invariant Recognition of Objects by Vision. *CBCL-291*, 2010.

[16] J. 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.

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

[18] G. Palm. On associative memory. *Biological Cybernetics*, 36:1931, 1980.

[19] 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.

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

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

[22] 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.

[23] R. P. Rao and R. D. L. Learning lie groups for invariant visual perception. 1999.

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

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

[26] 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.

[27] 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.

[28] 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.

[29] S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. Mathematics of the neural response. *Foundations of Computational Mathematics*, 10(1):67–91, 2010.

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

[31] G. Turrigiano and S. Nelson. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, (5):97–107, 2004.
[32] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 992–1006, 1991.
[33] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.
[34] G. Yu and J.-M. Morel. ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line*, 2011.

7. APPENDICES

7.1. **Appendix: Background.** In one of the early papers [25] we wrote:

It has often been said that the central issue in object recognition is the specificity-invariance trade-off: Recognition must be able to finely discriminate between different objects or object classes while at the same time be tolerant to object transformations such as scaling, translation, illumination, viewpoint changes, non-rigid transformations (such as a change of facial expression) and, for the case of cate- gorization, also to shape variations within a class.
and also

An interesting and non-trivial conjecture (supported by several experiments, of this population-based representation is that it should be capable of generalizing from a single view of a new object belonging to a class of objects sharing a common 3D structure such as a specific face to other views with a higher performance than for other object classes whose members have very different 3D structure, such as the paperclip objects. In a way very similar to identification, a categorization module say, for dogs vs. cats uses as inputs the activities of a number of cells tuned to various animals, with weights set during learning so that the unit responds differently to animals from different classes.

In the supermemo [26] I wrote:

Various lines of evidence suggest that visual experience – during and after development – together with genetic factors determine the connectivity and functional properties of units. In the theory we assume that learning plays a key role in determining the wiring and the synaptic weights for the S and the C layers. More specifically, we assume that the tuning properties of simple units – at various levels in the hierarchy – correspond to learning combinations of "features" that appear most frequently in images. This is roughly equivalent to learning a dictionary of patterns that appear with high probability. The wiring of complex units on the other hand would reflect learning from visual experience to associate frequent transformations in time – such as translation and scale – of specific complex features coded by simple units. Thus learning at the S and C level is effectively **learning correlations** *present in the visual world. The S layers' wiring depends on learning correlations of features in the image at the* **same time***; the C layers' wiring reflects learning correlations* **across time***. Thus the tuning of simple units arises from learning correlations in space (for S1 units the bar-like arrangements of LGN inputs, for S2 units more complex arrangements of bar-like subunits, etc). The connectivity of complex units arises from learning correlations over time, eg that simple units with the same orientation and neighboring locations should be wired together in a complex unit because often such a pattern changes smoothly in time (eg under translation).*

Since then we mainly focused on the hierarchical features represented by simple cells, on how to learn them from natural images and on their role in recognition performance. Here we focus on invariance and complex cells and how to learn their wiring, eg the domain of pooling.

As a consequence of this study, I have come to believe that I was wrong in thinking (implicitly) of invariance and selectivity as problems at the same level of importance. I now believe that the equivalence classes represented by complex cells are the key to recognition in the ventral stream. Learning them is equivalent to learning invariances and invariances are the crux of recognition in vision (and in other sensory modalities). I believe that the reason for multiple layers in the hierarchical architecture is the natural stratification of different types of invariances emerging from the unsupervised learning of the natural visual world with receptive fields of increasing size. In addition, the theory of this paper suggests that the tuning of the receptive fields in the hierarchy of visual areas depends in part from the transformations represented and discounted in each area.

7.2. **Appendix: Invariance and Templatebooks.**

- In applications if the templates are appropriately chosen the vectors s will also be made to be binary, and sparse enough to be close to optimal for associative retrieval of information (see section 6.5.1).

- Let us assume the classical case in which the templates are a set of $\phi(x)$. For reconstruction I would need a set such that $f(x) = \sum b_i \phi_i(x)$. Notice that a transformation of f is equivalent to transforming the templates ϕ, that is

$$\Pi f(x) = \sum b_i \Pi \phi(x).$$

  For the purpose of recognition instead of using the standard aggregation function $\Gamma = \sum$, it is possible to use a different aggregation function. For instance, in the case of piecewise constant representation (NN representation or vector quantization) I can use the compressed signature Σ_c given by

$$\Sigma_c = \text{argmax}_\phi \phi_i(x) \cdot f(x).$$

  The result Σ_c corresponds to the template ϕ_i, which is most similar to f.
- Normalization in the normalized dot product is needed to ensure invariance for scaling.

7.3. **Appendix: Affine Transformations in $\mathbb{R}^2$.** Let us assume – in this section – an x, y representation of images and transformations on them. In this representation, the components of the vector are the x, y coordinates of different features in an image. The features could be individual pixels *set in correspondence* across different images. A different representation that we will use in other parts of the paper is implicit and binary: we use 1 if a pixel is on and 0 otherwise. In this latter case, a vector corresponding to an image when displayed as a 2-D array is a binary image.

For each feature with coordinates x, y, we consider affine transformations defined as a 2×2 matrix

$$\Pi = \begin{pmatrix} a & b \\ d & e \end{pmatrix}. \tag{26}$$

Then an affine transformation with translations is

$$x' = \Pi x + t \tag{27}$$

with

$$x = \begin{pmatrix} x \\ y \end{pmatrix}$$

and

$$t = \begin{pmatrix} t_x \\ t_y \end{pmatrix}.$$

For a rotation of an angle θ the matrix $\mathbf{\Pi}$ is

$$\mathbf{\Pi} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}.$$

It is possible to represent in a more compact way affine transformations (including translations) using homogeneus coordinates with the vector $\mathbf{v}$

$$x = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

and the 3×3 matrix $\mathbf{\Pi_{A'}}$ acting on it

$$\mathbf{\Pi}' = \begin{pmatrix} a & b & t_x \\ d & e & t_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{28}$$

Thus $x' = \mathbf{\Pi}' x$.

Notice that the matrices $\mathbf{\Pi}'$ are representations of the affine group $Aff(2, \mathbb{R})$ which is an extension of $GL(2, \mathbb{R})$ by the group of translations in $\mathbb{R}^2$. It can be written as a semidirect product: $Aff(2, \mathbb{R}) = GL(2, \mathbb{R}) \times \mathbb{R}^2$ where $GL(2, \mathbb{R})$ acts on $\mathbb{R}^2$ in the natural manner.

7.3.1. *Decomposition of affine transformations.* There is another related decomposition of affine transformations, called the RQ decomposition. An homogeneous matrix A' can be decomposed as

$$A' = MK$$

where M is an orthogonal rotation matrix and $K = LS$ is an upper triangular matrix, L is a translation matrix and S is a shear and scale matrix. Thus

$$M = \begin{pmatrix} cos\theta & sin\theta & 0 \\ -sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$L = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix}$$

$$S = \begin{pmatrix} s_x & k & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$K = \begin{pmatrix} s_x & k & t_x \\ 0 & s_y & t_y \\ 0 & 0 & 1 \end{pmatrix}$$

.

7.3.2. *Approximation of general transformations in* $\mathbb{R}^2$. Assume that there are several features in the image (in the limit these features may be pixels in correspondence). Then the image can be represented as a vector

$$\begin{pmatrix} x_1 \\ y_1 \\ 1 \\ x_2 \\ y_2 \\ 1 \\ \cdots \\ x_N \\ y_N \\ 1 \end{pmatrix} \tag{29}$$

Assume the same affine transformation is applied to the whole vector. Then Π is

$$\mathbf{\Pi} = \begin{pmatrix} A & 0 & \cdots & 0 \\ 0 & B & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & Z \end{pmatrix} \tag{30}$$

where $A, B, \cdots, Z$ have the form of equation 9. If the same affine transformation is applied everywhere then the 2×2 blocks is such that $A = B = Z$ (this is the case we call *globally affine*).

7.4. **Appendix: Stratification.** .

- The size of an aperture is measured in terms of the wavelength of the highest spatial frequency contained in the visible patterns. In the case of V1 the input is from the LGN and can be represented as DOG filtered image. Notice that the optics of the human eye is bandlimited with an upper cut-off spatial frequency of 60 cycles per degree.
- Local linear approximations can approximate arbitrarily well a globally nonlinear function. This is what happens here with the approximation of transformations. The complexity of the transformation to be learned (linear vs strongly nonlinear) also affect sample complexity: much fewer data are needed to learn a local affine transformation than a complex global one. This may represent an argument to show why a hierarchical architecture – with increasing aperture sizes (eg decreasing number of apertures) and limited to learn simple affine transformations in each aperture – is optimal or suboptimal.
- The Jacobian determinant at a given point gives important information about the behavior of a function T near that point. For instance, the continuously differentiable function T is invertible near a point $\mathbf{x}_0$ if the Jacobian determinant at the point is non-zero. This is the inverse function theorem. The absolute value of the Jacobian determinant at $\mathbf{x}_0$ gives us the factor by which the function T expands or shrinks volumes near $\mathbf{x}_0$.
- In lemma 3 if we assume that the image is a smooth function of x, y we can then represent it up to linear terms within a small enough patch in terms of its Taylor series around the center of the patch. Then the proof can use a Taylor expansion of the image around a point x_0, y_0 within a patch. The assumption that approximation error with linear and constant terms should remain $\leq K$ then determines the size of the patch.
- A radially symmetric aperture (a disk) corresponds to convolution in the Fourier domain with $J_1(\omega r)$, where J_1 is the Bessel function of the first kind. A large aperture corresponds to an increasingly delta-like Fourier transform; a small aperture corresponds to broader and broader envelope and more and more "blurring" (in the frequency domain). The same reasoning can be repeated with a radially symmetric spatial Gaussian modeling the "aperture". It is also important to remember that we consider affine transformations which are *uniform* within an aperture (eg within the receptive field of one cell). Clearly, a large range of complex, non-uniform *global* transformations of the image can be approximated by local affine transformations.
- Statistics of image transformations: Hinton mentions that even at a fairly large patch size, uniform shifts of the entire visual field are the predominant mode of variability in broadcast video.
- The usual representation of an affine transformation in $\mathbb{R}^2$ is such that the sequence of transformations is rotation and scaling followed by translation. This implies that the inverse of it – which is the transformation of interest to us – is translation followed by rotation+scaling. For uniform scaling the order of rotation and scaling does not matter since they commute. For inverse nonuniform scaling a possible decomposition (after translation) is rotation followed by nonuniform scaling followed by another rotation.

- Do the statistics of natural (image) transformations (how frequent rotations of different types are, translations etc) play a role?
- *Stratification conjecture* Translation invariance is achieved by C units which pool together the responses to the same set of S templates at many different locations. The association can be performed on the basis of the temporal continuity of the S templates activated by the object translating. The S templates could be pre-existing or learned simultaneously with their association by temporal contiguity (see later). Any transformation could be learned by simple association of templates memorized during the transformation of an object.

7.5. **Appendix: Spectral Properties of the Templatebook.**

7.5.1. *Spectral Properties of the Translation Operator.* The eigenfunctions depend on the representation we use for images. The standard representation is in terms of x, y coordinates of corresponding features or points in images. In this explicit representation of images as vectors of x, y coordinates, translations cannot be mapped to matrices acting on the vectors, unless I use homogeneous coordinates (see section 7.3). As we will see, in this representation translations do not commute with scaling and rotation; scaling and rotation commute which other only if the scaling is uniform.

It is well known that the eigenfunctions associated with the translation operator (in R^2 in our case) are the complex exponentials. The informal argument runs as follows. Consider the translation operator acting on functions $\phi(x)$ in L_2 defined by $T_{x_0}\phi(x) = \phi(x - x_0)$. The operator $T_{x_0} = e^{-x_0 \frac{d}{dx}}$ is unitary and forms a representation of the additive group. The definition leads to a functional eigenvalue equation

$$\phi(x - x_0) = \lambda\phi(x)$$

with solutions (see Supp. Mat. [21]) $\phi(x) = e^{i\omega x}$.

7.5.2. *Spectral properties of the uniform scaling and rotation operators.* The eigenfunctions of rotations and uniform scaling are complex exponentials in polar coordinates. In other words $\phi(x, y) = \rho e^{i\theta}$ is a solution of the eigenvalue equation for the rotation operator R

$$R_{\psi_0}\phi = \lambda\phi$$

with $\lambda = e^{i\psi_0}$, and similarly for the scaling operator, where the eigenvalue is real.

7.5.3. *Compositions of transformations.* Assume the semidirect product $Aff(2, \mathbb{R}) = GL(2, \mathbb{R}) \times \mathbb{R}^2$ for a composite transformation that I have introduced earlier. Let us focus on the linear transformations represented by a two-by-two matrix A, neglecting translations. A can be decomposed using SVD as

$$A = U\Sigma V^T$$

where all matrices are 2×2, Σ is diagonal and U and V are orthogonal. Thus any affine transformation represented in this way is decomposed into a rotation followed by asymmetric scaling followed by a rotation. It follows that the condition number of A is 1 if scaling is isotropic and larger than 1 otherwise. It is possible to consider a sequence of transformations such as for instance scaling and rotation and analyze it in terms of the SVD decomposition.

7.5.4. *SVD of a "movie": temporal order (by J. Mutch).* The typical SVD setup is

A is $M \times N$ matrix, $A = USV^T$, where U is $M \times M$, S is $M \times N$, and V is $N \times N$. Now suppose I permute the columns of A. Then:

- The matrix U is unchanged, except that some of the columns might get multiplied by -1.
- The matrix S is unchanged.
- The matrix V is different.

Thus, SVD depends only on the entire "cloud" of frames we give it and temporal ordering is irrelevant.

7.5.5. *Gabor frames.* The windowed Fourier transform (WFT) and the inverse are

$$F(\omega, a) = \int dx f(x) G(x-a) e^{-i\omega x} \tag{31}$$

$$f(x) = \frac{1}{||G^2||} \int d\omega da G(x-a) F(\omega, a) e^{i\omega x} \tag{32}$$

An examination of the first equation shows that $F(\omega, a)$ is the Fourier transform of $f(x)G(x-a)$, that is *the pattern $f(x)$ "looked at" through a Gaussian window $G(x)$ centered at a.* Since Fourier components emerge from translation, this implies that Gabor wavelets of the form $G(x)e^{-i\omega x}$ emerge from translations of $f(x)$ modulated by $G(x)$.

The previous argument is for a single (Gaussian, say) fixed aperture centered in a. One can ask whether $f(x)$ can be represented in terms of several apertures spanning a lattice in x, ω space. The answer is affirmative for $x \in \mathbb{R}^1$ (see [4] and references therein): the Gabor system (composed of Gabor or Weyl-Heisenberg frames)

$$\mathcal{G}_\Lambda = \{e^{2i\pi yt} e^{-\pi(t-x)^2} : (x, y) \in \Lambda\} \tag{33}$$

is a frame for a sequence Λ of sufficient density.

7.5.6. *Gabor wavelets.* Separately, it has been argued that actual Gabor wavelets (with σ depending on ω) are a representation of the similarity group (in R^2). Stevens [30] develops an interesting and detailed argument for Gabor receptive fields in V1 to be implied by "invariance" to translations, scale and rotations. His Gabor wavelets have the form

$$\phi_{\theta,\xi,\eta}(x, y) = e^{-\frac{(x_\theta - \xi_\theta)^2}{2\sigma_x^2}} e^{-\frac{(y_\theta - \eta_\theta)^2}{2\sigma_y^2}} e^{i2\pi x_\theta \omega}$$

where the center of the receptive field is ξ_θ, η_θ, the preferred orientation is θ and σ_x^2 and σ_y^2 are proportional to $\frac{1}{\omega^2}$.

This family of 2D wavelets, and their 2D Fourier transforms, is each closed under the transformation groups of dilations, translations, rotations, and convolutions.

7.5.7. *Gabor frames diagonalize the templatebooks acquired under translation through a Gaussian window.* The argument is about the spectral content of a row of the templatebook acquired by recording frames of a video of an image patch translating and looked at through a Gaussian window. The row of the templatebook is a tensor with the rowindex denoting time and the other two indeces running through the x and y coordinates of the image patch. Suppose that the translation is parallel to the x coordinate. Then "flattening" the tensor gives a Toeplitz matrix, since matrices at time i and $i+1$ are shifted along x. Toepliz matrices have spectral properties closely related to circulant matrices. For simplicity I describe here, as an example, the analysis for circulant matrices. This is what I would obtain from each row of the template book if the patterns moving behind the Gaussian window would be periodic (eg have the geometry of a thorus). In this case, the DFT diagonalizes the circulant matrix X. Thus

$$F^T X F = \Lambda,$$

where Λ is a diagonal matrix. In particular, a column of the matrix, which is an image "looked at" through a Gaussian window, can be represented as $GI = G \sum c_l e^{i\omega l x} = \sum c_l G e^{i\omega l x}$ thus in terms of Gabor frames.

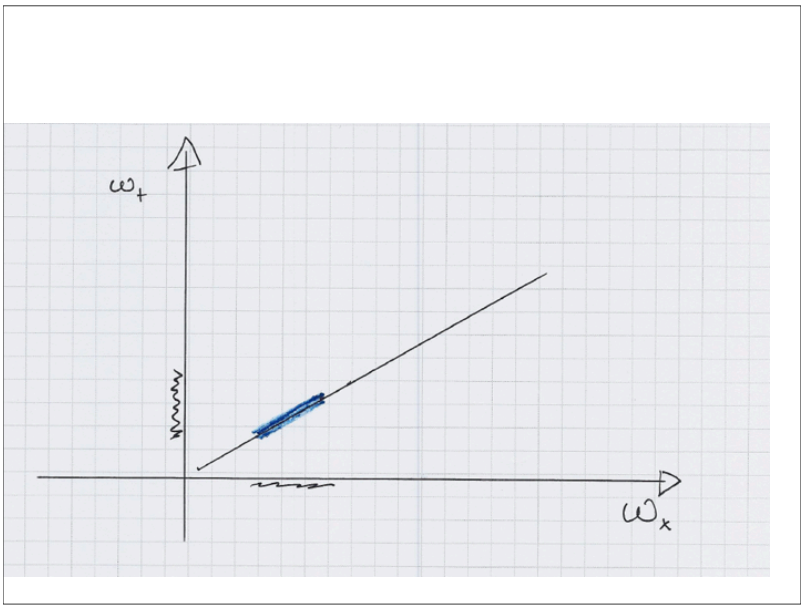


FIGURE 27. *For a object moving at constant speed in 1D the support of the spatial spectrum is on the line corresponding to $v = const$. Temporal bandpass filtering affects spatial frequency (see [1]). The spatial 2D case deserves some additional thought..*

Well-know results (see for instance [4]) extend considerably the math of this section – for instance providing conditions on the lattice of the "apertures" to ensure good global representations. It would be interesting to explore this issue and the related one about sampling the scale space and which mechanisms during development may be responsible for it$\odot$.

7.5.8. *Lie algebra and Lie group.* The Lie algebra associated with the Lie group of affine transformations in $\mathbb{R}^2$ has as an underlying vector space the tangent space at the identity element. For matrices A the exponential map takes the Lie algebra of the general linear group G into G.

Thus a transformation T of x parametrized by t can be represented as $T = e^{At}$. Notice that if A is symmetric then $A = U\Lambda U^T$ and

$$T = e^{At} = e^{U\Lambda U^T t} = U e^{\Lambda t} U^T. \tag{34}$$

and thus *the spectrum of A and the spectrum of T coincide*. This is not true if A is not symmetric.

7.5.9. *Oja's flow.* An interesting version of Oja's rule is the version that applies to D' neural units – in our case the D' complex cells associated with the first rows of a templatebook:

$$\mathbb{W}_{k+1} = \mathbb{W}_k + \mu_k[x_k - \mathbb{W}_k y_k]y_k^T \tag{35}$$

where $\mathbb{W}_{k+1} = [w_k(1)w_k(2)\cdots w_k(m)]$ is the weight matrix whose columns are the individual neuron weight vectors $w_k(i)$ and $y_k = \mathbb{W}_k^T x_k$ is the output vector of D elements.

7.6. **Appendix: Mathematics of the Invariant Neural Response (with L. Rosasco).** We provide a mathematical description of the neural response architecture [29] of Figure 1, which is designed to be robust to transformations encoded implicitly in sets of templates. Robustness is achieved by mean of suitable pooling operations across the responses to such templates. The setting we describe is a modification of the one introduced in [29].

7.6.1. *Framework.* We start giving the basic concepts and notations describing the framework we consider.

Architecture Elements The new neural response is defined by an architecture composed of the following elements.

- A finite number of nested sets $p_1 \subset p_2 \subset \cdots \subset p_n$, that we call patches.
- A family of function spaces defined on each patch

$$(\mathrm{Im}(p_i))_{i=1}^n, \quad \text{where} \quad \mathrm{Im}(p_i) = \{x \mid x : p_i \to [0,1]\}, \quad i = 1, \ldots, n.$$

- A family of finite sets of maps from a patch to the next larger one,

$$(H_i)_{i=1}^{n-1}, \quad \text{where} \quad H_i = \{h \mid h : p_i \to p_{i+1}\},$$

that we call decomposition maps. The name is justified by the observation that H_i describe how a function $x \in \mathrm{Im}(p_{i+1})$ can be decomposed in a set of functions $x \circ h$, $h \in H_i$, with smaller domain, namely a set of *parts*.

7.6.2. *Tuning Function.* A tuning function $K : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ is given, which is a reproducing kernel Hilbert space. The tuning function can be naturally restricted to $\mathbb{R}^b \times \mathbb{R}^b$, with $b \leq d$. The two main examples of tuning function we have in mind are the Gaussian $K(x,x') = \exp -\gamma \|x - x'\|^2$ and the normalized inner product $K(x,x') = \frac{\langle x,x' \rangle}{\|x\|\|x'\|}$, where $\langle \cdot, \cdot \rangle$, $\|\cdot\|$. are the inner product and norm in $\mathbb{R}^d$.

7.6.3. *Families of Invariance Sets.* A last crucial ingredient is needed to define the generalized neural response. A family of sets whose elements are themselves sets of functions, that is

$$(V_i)_{2=1}^n \quad \text{where} \quad V_i = \{v \mid v = \{t \mid t \in \mathrm{Im}(p_i)\}\}, \quad i = 2, \ldots, n.$$

We assume that $|V_i| \leq d$, and $|v| \leq d$ for $v \in V_i$ and all $i = 2, \ldots, n$. Each element v of a set V is called an invariance set.

7.6.4. *New Neural Response Definition.* The definition of the generalized invariant neural responses is the following.

Definition 5. *Given an initial neural response $N_1 : \mathrm{Im}(p_1) \to \mathbb{R}^p$, $p \leq d$, the m-layer neural response $N_m : \mathrm{Im}(p_m) \to \mathbb{R}^{|V_m|}$, for $m = 2, \ldots, n$, is defined as*

$$N_m(x)(v) = \max_{t \in v} \left\{ \sum_{h \in H} K(N_{m-1}(x \circ h), N_{m-1}(t \circ h)) \right\} \tag{36}$$

with $x \in \mathrm{Im}(p_m)$, $h \in H_{m-1}$, $v \in V_m$.

7.6.5. *Learning.* The interpretation of the above model that suggests how the invariance can be learned from data. Having in mind problems in computational vision, we think of $\mathrm{Im}(p_i)_{i\geq 1}$ as images of increasing size.

The patches can be thought of as squared domains centered around the origin. The decomposition maps describe how an image can be decomposed into (possibly overlapping) image patches. In the above construction an invariance set $v \in V_m$ is often an ordered set of images. In practice this set can be obtained from a video sequence v so that, if $v = \{t_1, \ldots, t_p\}$ then $t_1, \ldots, t_p$ correspond to frames at successive instants of time. The recording of sets of video sequences is the learning phase of the above model.

7.7. **Restricted Appendix: Future projects, open questions and garbage collection.**

7.7.1. *Definitions and Theorems we need to formulate (and in a couple of cases to prove).*

Definition 6. *A set of transformations... on images...seen through a Gaussian aperture*

Definition 7. *The group of affine transformations consists of rotations scalings and translations. The affine group $Aff(2,\mathbb{R})$ which is an extension of $GL(2,\mathbb{R})$ by the group of translations in $\mathbb{R}^2$.*

Theorem 7. *$Aff(2,\mathbb{R})$ can be written as a semidirect product: $Aff(2,\mathbb{R}) = GL(2,\mathbb{R}) \times \mathbb{R}^2$ where $GL(2,\mathbb{R})$ acts on $\mathbb{R}^2$ in the natural manner. Each of the subgroups (transla, rotation, scaling) is LCA. Rotation and (symmetric?) scaling commute and have same characters but translations and the rest does not.*

Lemma 5. *A representations of the affine group $Aff(2,\mathbb{R})$ is given by the matrices $\mathbf{\Pi}'$.*

Lemma 6. *The generators of the Lie algebras corresponding to the subgroups of the affine group are...*

Definition 8. *A templatebook is generated by $g^n t_0$*

Theorem 8. *For discrete ? affine subgroups by averaging over the subgroup (ie $R_G[f(x)] = \frac{1}{|G|}\sum_{g\in G} f(g(x))$) the following aggregation function on a patch $I(x)$ gives a number which is invariant to any transformation of I: $R_G[I(x)] = \frac{1}{|G|}\sum_{g\in G} |\int I(x) g\circ\chi(x)dx|^2 = \frac{1}{|G|}\sum_{g\in G} |\tilde{I}(\omega)|^2 = |\tilde{I}(\omega)|^2$*

Lemma 7. *The characters of each of the subgroups of Aff are Fourier $\chi(x) = e^{i\omega x}$and any function ? can be represented in terms of them*

Theorem 9. *Empirical eigenvectors of templatebook converge ??? to $G(x - x_0)e^{i\omega x}$*

Theorem 10. *For each subgroup $|\tilde{I}(\omega)|^2$ is invariant*

Theorem 11. *The Oja flow of sample rows of the templatebook converges to the character of the group*

Theorem 12. *Stratification. Suppose a finite large window (visual field). Suppose that rotations, expansions and trans are happening with same probability and random origin. Suppose an aperture in random position. For aperture size decreasing ...for most apertures the best estimate (for a given amount of noise) is translation*

Questions
what are the main PCAs of faces rotating in depth? Simulations...
what does Oja flow converge to?
should we work with the quotient group from layer to layer (peeling off)? Theorem?

7.7.2. *Visual abstractions.*

- *Line drawings* Suppose there are templatebooks that contain equivalence sets for grey level edges and drawings. Then line drawing may be quasi-equivalent to natural images. This has to be formulated and looked at in detail$^{\oplus}$. **Get cartoons and affine transformations of them from Andrea!** Consider adelson normal from learning from line drawings of complex curved shapes associating different line drawings together (under transformation)...this can be done even if line drawings are not in the object but in the image and are ridges of luminance$^{\oplus}$.
- *Collinearity* Similarly to the point about parallel lines: one example is enough to cover a large number of instances of collinearity$^{\oplus}$. Are cells in higher areas (V2?) sensitive/invariant to collinearity?
- *Triangles* Same as above because of invariance to affine transformations
- *Numerosity* Consider a small number, say 3, of squares. Consider all affine transforms and possibly blrurring as well as derivative high-frequency filtered versions. This will consist a large set of images of 3 elements.
- *Bilateral or other symmetry* yes, if representationbelow is closed and complete up to that layer – then it is just a property of the last templates (bilateral symmetry or any other symmetry)
- *Invariance does not mean blindness* The architecture provides information to stages above it (such as a classifier) from all areas. This allows to have for instance selectivity to object identity while invariant to position but at the same time some selectivity to position independent from identity (as it is experimentally the case in IT cells).
- *Compositionality* Is it possible and does it make sense to have templates that compose patterns in more complex ones? I am thinking of a horse and a rider. The templates would involve a transformation from a person to a person riding a horse. They may allow an approximative invariance in recognizing a person as a person independently from whether the person is riding an animal$^{\oplus}$.
- *Classes of equivalence and concepts* Classes of equivalence are pretty abstract, so this is a promising **project**.
- consider transformation of programs (paper on using genetic algs)

7.7.3. *Against a "naive slowness" principle.* A naive slowness principle will prefer features that are parallel to the motion underlying the transformation (lines parallel to the direction of translation, circles for rotations, star patterns for expansion) whereas the approach here chooses exactly the orthogonal features (lines orthogonal to the direction of translation, star patterns for rotations, circles for expansion) and then builds invariance to their shifts. The slowness principle of Maurer (who has a cleaner formulation than Wiskott) may not have the full extent of this problem because it adds to slowness a constraint of maximal variance of the features. Wiskott's Figure 12 is telling. For rotation, for example, they get ring-like features, i.e. features that are parallel to the direction of the transformation, not orthogonal.

Notice that spatial Fourier components in the image which are parallel to the direction of motion are effectively filtered out at the level of the LGN by temporal bandpass filtering.

7.7.4. *Invariances and constraints$^{\oplus}$.* Symmetries and invariances are almost equivalent (in the language of physics). Some of the interesting constraints are implied by general invariances, such as invariance to the choice of units of measure. They also correspond to constraints (for instance consider $E - mc^2 = 0$). Appropriate constraints are key in making an optimization problem well-posed and in particular a learning algorithm to be predictive.

7.7.5. *Notes and Questions.*

- Suppose that the normalized dot product is replaced by a dot product. Suppose that the max is replaced by the average. Will the hierarchical architecture still work? If yes, this would be somewhat surprising because this simplified architecture is completely linear. A possible reason may ve that hierarchical architectures may be decomposing a nonlinear problem in locally linear problems.
- *Analysis of linearized architecture:* it should work even if with – probably – suboptimal discriminability. The reason to do it is that the analysis should be easy.
- *HMAX with average:* It should be tried again: how does it do on standard databases?
- There is a possible "generative" version of the model for learning invariance to transformations outlined in the main text. Each layer may transform the input "image" – instead of simply encoding invariant signatures. In this case the transformation operator – a matrix – could be learned by an associative memory module which associates a templateset $x = \tau$ with its transformed version $y = T\tau$, both observed as frames of a "video" recorded during a transformation. Thus T represents the transformation between x and y and can transform x into y.

  The (non-symmetric) matrix K is learned by associating a templateset $x = \tau$ to its transformed version $y = T\tau$, both observed as frames of a "video" recorded during a transformation. Let us assume that there are many, say M, such pairs, thus M columns of the matrix X and of the matrix Y[9]. For simplicity we assume K to be optimal in the least-square sense. Thus $K = YX^\dagger = YX^T(XX^T)^{-1}$. We are interested in the structure of K as it is learned layer after layer.

  Gedankenexperiment: learning transformations and receptive fields using RDI One of the underlying hypotheses here is that the receptive fields in the various areas of the ventral stream are determined by the transformations represented in each area. It may be interesting for empirical studies to use feature-free patterns in order to learn transformations such as RDIs – eg Random Dot Images. If the hypothesis is right, interesting receptive fields should emerge. Notice that for "implicit" representations of RDIs in terms of binary arrays $X^\dagger$ is $X^\dagger \sim I$ and thus $K \sim YX^T$, where YX^T is the empirical correlation function of x and y. In a sense, K is a *generative model* for whatever discriminative scheme uses the associated information. The spectrum of XX^T and YX^T are similar in this framework. It follows that there there is a very intimate relationship between the learned K and Glass [3] patterns, which are superpositions of two RDIs, eg x and y.

  Notice that so far I have used vectors for images: for instance x_i is a component of the vector x representing a 2D image. Thus the index i is equivalent to some k, l (with $k, l \to i$) in the array representation of the image.

  This Gedankenexperiment (also done in reality) suggest that the tensor K when trained with translations in a fixed direction, will translate pixels in the array x in a direction in the plane, that is $K_{i,j,k,l} \sim \delta(i+1,k)\delta(j,l)$ for horizontal translations. When trained with rotations will *rotate* pixels.
- Let us assume that T_{x_0} is represented by a linear kernel $T_{x_0}(x, y)$. Assume that its action on $\phi(x)$ is described by

$$\int T_{x_0}(x-u)\phi(u) = \phi(x-x_0)$$

  and in addition

$$\int T_{x_0}(x-u)\phi(u-u_0) = \phi(x-x_0-u_0)$$

[9]The rows of $\mathbb{T}$ are now columns in X and Y.

for any u_0, to reflect natural properties of translation. Because of linearity and shift-invariance the Fourier transform "diagonalizes" T giving

$$\hat{T_{x_0}}\hat{\phi} = \hat{\phi}e^{-i\omega x_0}$$

from which we derive

$$\hat{T_{x_0}} = e^{-i\omega x_0}$$

.
and $T_{x_0}\phi(x) = \phi(x + x_0)$

- Spectrum of small apertures reflects transformations independent of objects; large apertures spectrum reflect class. Notice that the size of the aperture controls the complexity eg the degrees of freedom of images there.
- In small apertures learning may have a long time constant, thus what is learned is across many different objects and becomes independent of objects. In large apertures time constant may be short so only indivisual videos are stored and learned in order to have class-specific invariances.
- The fully linear case may be a good simple example to think about: averaging and convolutions and subsampling from layer to layer.
- is normalized dot product effectively linear? If yes and if max is replaced by average then where are nonlinearities? could a fully linear architecture be enough? Notice that average preserve invariance of signatures from layer to layer...However, it may loose in resolution so max or other nonlinear operation may be better. A good example is given by zero-crossings
- Question: WHY from V1 to V2 to V4 to IT a factor 2 per step in RF size? Is it because of correpsondence? Doubling in α slope in $size = \alpha eccentricity$ (roughly) from V1 to V2 to V4: check gadass? Is because of increase in S and C?
- Body pose areas—thus animate+inanimate? The distinction could also be because of self-motion vs no self-motion...
- Place area may involve perspective transformations
- Estimate signature invariant of pose AND pose invariant of identity (see above). Is there a role for max over pose and max over identity (like in supervector?)?
- which invariances etc for dorsal stream?
- is the dorsal stream evolutionary older? what about development?
- Not reconstruction. Just signature More in general if goal is not reconstruction but matching then like stereo I want large neighborhood with complex features and small N with simple features...important to have dictionaris which are not too big i each N and to avoid false target problem. Each N has its dictionary of t and of Tt...
- We conjectured (see Supermemo) that generic as well as class-specific transformations are learned by exploiting correlations in time (via mechanisms such as trace rule)
- Hmax can learn (unsupervised, from visual experience) invariance to generic transformations eg translations (cite Masquellier serre poggio) and invariance to class-specific transformations eg viewpoint. The model hierarchy figure shows that C3 features do well on viewpoint and the other figure shows thatC3 features may be class-specific (the figure that shows the results of encoding templates on the wrong class)
- S cells (features) are not so critical (see results from other groups e.g. Serre, LeCun, Ng, Leibo)
- C cells (invariances/ classes of equivalence) are key
- we conjecture that in evolution the hierarchy of cortex follows from the need to learn from experience increasingly complex transformations from translation and scales to viewpoint and body pose

- The visual cortex inherit symmetry properties of the physical world, in part through evolution and in part through early visual experience during development, through hierarchies of simple *associative learning* mechanisms.
- a neuron with 3 bits in the model is equivalent to 3 binary neurons: in which sense? See memo of Sharat...Notice that in general a vector of n elements can be encoded as a number say binary
- Under orthographic projection 3D affine reduces to GL(2), probably nonlinear subspace of 2D affine.
  - Glass patterns experiments at various eccentricities
  - adapt an orientation; invariance to small rotations locally orthogonal to it should be affected; same (but orthogonal) for expansions;
- One of the main predictions is that visual areas are involved in representing different geometric transformations. Furthermore the *rough* sequence should be: V1 for translations, V2 for translations and scale and rotations, V4 for affine (mainly scale). The idea is to use Glass patterns and have subjects look at them through an aperture (a circle). First in the fovea. I expect that if the aperture is too small nothing can be seen. At some point translations will be detectable but not rotations or scaling. Making the aperture larger should allow detection and discrimination of rotations and scalings (not sure which one thirst). First we should check this in the fovea. If this works, then we should try the same threshold discrimination experiment at a couple of eccentricities and plot, as a function of eccentricity, the minimum aperture for which translations and rotations and scaling can be first detected. The ideal figure is enclosed (we assume we know how receptive field size grows from V1 to V2 etc as a function of eccentricity). This of course is a dream (it would suggest that areas are involved in transformation and localize which transformation in which area). Even if true it is not clear that V1, v2 etc are the bottleneck for the psychophysics (conscious perception?). This is a big extra assumption that has little to do with the theory.

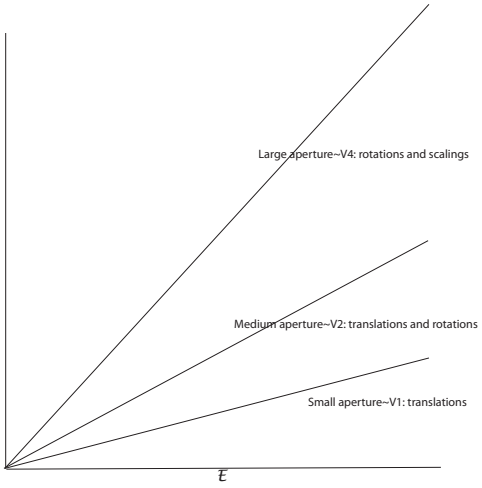


FIGURE 28. Conjecture on threshold perception of type of transformation for Glass patterns as a function of eccentricity and aperture.