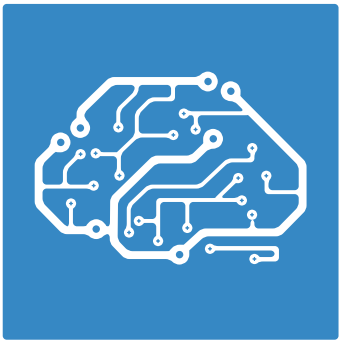

Center for Brains, Minds & Machines

CBMM Memo No. 026 **November 14, 2014**

Representation Learning in Sensory Cortex: a theory

by

Fabio Anselmi[⋆,†] **and Tomaso Armando Poggio**[⋆,†]
⋆ **Center for Brains, Minds and Machines,**
McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA
† **Istituto Italiano di Tecnologia, Genova, Italy**

Abstract: We review and apply a computational theory of the feedforward path of the ventral stream in visual cortex based on the hypothesis that its main function is the encoding of invariant representations of images. A key justification of the theory is provided by a theorem linking invariant representations to small sample complexity for recognition - that is, invariant representations allows learning from very few labeled examples. The theory characterizes how an algorithm that can be implemented by a set of "simple" and "complex" cells - a "HW module" – provides invariant and selective representations. The invariance can be learned in an unsupervised way from observed transformations. Theorems show that invariance implies several properties of the ventral stream organization, including the eccentricity dependent lattice of units in the retina and in V1, and the tuning of its neurons. The theory requires two stages of processing: the first, consisting of retinotopic visual areas such as V1, V2 and V4 with generic neuronal tuning, leads to representations that are invariant to translation and scaling; the second, consisting of modules in IT, with class- and object-specific tuning, provides a representation for recognition with approximate invariance to class specific transformations, such as pose (of a body, of a face) and expression. In the theory the ventral stream main function is the unsupervised learning of "good" representations that reduce the sample complexity of the final supervised learning stage.

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

Representation Learning in Sensory Cortex: a theory

November 19, 2014

Fabio Anselmi[⋆,†] and Tomaso Armando Poggio[⋆,†]
⋆ *Center for Brains, Minds and Machines, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA*
† Istituto Italiano di Tecnologia, Genova, Italy

Summary *We review and apply a computational theory of the feedforward path of the ventral stream in visual cortex based on the hypothesis that its main function is the encoding of invariant representations of images. A key justification of the theory is provided by a theorem linking invariant representations to small sample complexity for recognition - that is, invariant representations allows learning from very few labeled examples. The theory characterizes how an algorithm that can be implemented by a set of "simple" and "complex" cells - a "HW module" – provides invariant and selective representations. The invariance can be learned in an unsupervised way from observed transformations. Theorems show that invariance implies several properties of the ventral stream organization, including the eccentricity dependent lattice of units in the retina and in V1, and the tuning of its neurons. The theory requires two stages of processing: the first, consisting of retinotopic visual areas such as V1, V2 and V4 with generic neuronal tuning, leads to representations that are invariant to translation and scaling; the second, consisting of modules in IT, with class- and object-specific tuning, provides a representation for recognition with approximate invariance to class specific transformations, such as pose (of a body, of a face) and expression. In the theory the ventral stream main function is the unsupervised learning of "good" representations that reduce the sample complexity of the final supervised learning stage.*

1 Intro and background

The ventral visual stream is believed to underlie object recognition abilities in primates. Fifty years of modeling efforts, which started with the original

Hubel and Wiesel proposal of a hierarchical architecture iterating in different layers the motif of simple and complex cells in V1, led to a series of quantitative models from Fukushima 1980 to HMAX (Riesenhuber and Poggio 2000) which are increasingly faithful to the biological architecture and are able to mimic properties of cells in different visual areas while achieving human-like recognition performance under restricted conditions. Recently, deep learning convolutional networks which are hierarchical and similar to HMAX but otherwise do not respect the ventral stream architecture and physiology, have been trained with very large labeled datasets (Russakovsky et al. 2014, Google 2014, Zeiler and Fergus 2014). The population of resulting model neurons mimic well the object recognition performance of the macaque visual cortex (Dicarlo, unpublished). However, the nature of the computations carried out in the ventral stream is not explained by such models that can be simulated on a computer but remain otherwise rather opaque.

In other papers (in particular Anselmi et al., 2014; Poggio et al., 2014) we have developed a mathematics of invariance that can be applied to the ventral stream. In this neuroscience paper we do this and outline a comprehensive theory of the feedforward computation of invariant representations in the ventral stream (Anselmi et al. 2014, Anselmi et al. 2013) - that is a theory of the first 100 milliseconds of visual perception, from the onset of an image to activation of IT neurons about 100 msec later. In particular, such representations are likely to underlie rapid categorization – that is immediate object recognition from flashed images (Potter 1975, Thorpe et al 1996). We emphasize that the theory is not a full theory of vision that should explain top down effect and the role of backprojections, but only a precursor to it. The theory, dubbed i-theory, is based on the hypothesis that the main computational goal of the ventral stream is to compute neural representations of images that are invariant to transformations commonly encountered in the visual environment and learned from unsupervised experience. I-theory proposes computational explanations for various aspects of the ventral stream architecture and of its neurons. It makes several testable predictions. It also leads to network implementations that show high performance in object recognition benchmarks (Liao et al. 2013). As we mentioned, the theory is based on the unsupervised, automatic learning of invariant representations. Since invariant representations turn out to be "good" representation for supervised learning, characterized by small sample complexity, the architecture of the ventral stream may ultimately be dictated by the need to learn from very few labeled examples, similar to human learning but quite different from present supervised machine learning algorithms trained on large sets of labeled examples.

We use i-theory to compactly summarize several key aspects of the neuroscience of visual recognition, explain them and predict others. The organization of the paper is as follows. Section 2 and 3 describe the general theoretical framework: a computational theory of invariance in section 2 and a theory of the basic biophysical mechanisms and circuits in section 3. In particular, we describe relevant new mathematical results on invariant representations in vision that are given elsewhere with details and proofs (Anselmi et al. 2014, Anselmi

et al. 2013). The starting point is a theorem proving that image representations which are invariant to translation and scaling and approximately invariant to some other transformations (e.g. face expressions) can considerably reduce the sample complexity of learning. We then describe how an invariant and unique (selective) signature can be computed for each image or image patch: the invariance can be exact in the case of locally compact group transformations (we focus on groups such as the affine group in 2D and one of its subgroups, the similitude group consisting of translation and uniform scaling) and approximate under non-group transformations. A module performing filtering and pooling, like the simple and complex cells described by Hubel and Wiesel (HW module), can compute such estimates. Each HW module provides a feature vector, which we call a signature, for the part of the visual field that is inside its receptive field. Interestingly, Gabor functions turn out to be optimal templates for maximizing simultaneous invariance to translation and scale. Hierarchies of HW modules inherit their properties, while alleviating the problem of clutter in the recognition of wholes and parts. Finally, the same HW modules at high levels in the hierarchy are shown to be able to compute representations which are approximately invariant to a much broader range of transformations—such as 3D expression of a face, pose of a body, and viewpoint, by using templates, reflected in neuron's tuning, that are highly specific for each object class. Section 3 describes how neuronal circuits may implement the operation required by the HW algorithm. It specifically discusses new models of simple and complex cells in V1. It also introduces plausible biophysical mechanisms for tuning and pooling and for learning the wiring based on Hebbian-like unsupervised learning.

The rest of the paper is devoted to reviewing the application of the theory to the feed forward path of the ventral stream in primate visual cortex. Section 4 applies the theory to explain the multi resolution, eccentricity-dependent architecture of the retina and V1 as a consequence of the need for simultaneous space and scale invariance. It predicts several still untested properties of the early stages of vision. Section 5 deals with V2 and V4 as higher layers in the hierarchy devoted to progressively increase invariance to shift and scaling while minimizing interference from clutter ("minimizing crowding"). Section 6 is about the final IT stage where class-specific representations that are quasi-invariant to non-generic transformations are computed from a shift and scale invariant representation obtained from V4. It also discusses the modular organization of anterior IT in terms of the theory; in particular it proposes an explanation of the architecture and of some puzzling properties of the face patches system. We conclude with a discussion of predictions to be tested and other open problems.

2 Computational level: mathematics of invariance

As context for this paper, let us describe the conceptual framework for primate vision that we use:

- The first 100ms of vision in the ventral stream are mostly feed forward. The main computation goal is to generate a number of image representations each one or quasi invariant to some transformations experienced during development and at maturity, such as scaling, translation, and pose changes. The representations are used to answer basic default questions about what kind of image and what may be there.
- The answers will often have low confidence requiring an additional "verification/prediction step" which may require a sequence of shifts of gaze and attentional changes. This step may rely on generative models and probabilistic inference and/or on top-down visual routines following memory access. Routines that can be synthesized on demand as a function of the visual task are needed in any case to go beyond object classification. Notice that in a Turing test of vision (see http://cbmm.mit.edu/research-areas/cbmm-challenge/) only the simplest questions (what is there? who is there?...) can be answered by pretrained classifiers.

We consider only the feedforward architecture of the ventral stream and its computational function. To help the reader understand more easily the mathematics of this chapter, we anticipate here the network of visual areas that we propose for computing invariant representations for feedforward visual recognition. There are two main stages: the first one computes a representation that is invariant to affine transformations, followed by a second stage that computes approximate invariance to object specific, non-group transformations. The second stage consists of parallel pathways, each one for a different object class (see Figure 4 stage1). The theorems of this section do not strictly require these two stages: the second one may not be present, in which case the output of the first directly access memory or classification. If both are present, as it seems the case for the primate ventral stream, the mathematics of the theory requires that the object specific stage follows the one dealing with affine transformations. According to the theory, the HW module mentioned earlier is the basic module for both stages. The first and the second stage pathways may consist of a single layer of HW modules. However, mitigation of interference by clutter requires a hierarchy of layers (possibly corresponding to visual areas such as V1, V2, V4, PIT) within the first stage. It may not be required in visual systems with lower resolution such as the mouse. The final architecture is shown in the Figure 4: in the first stage about four layers compute representations that are increasingly invariant to translation and scale while in the second stage a large number of specific parallel pathways deal with approximate invariance to transformations that are specific for objects and object-classes. Notice that for any representation which invariant to X and selective for Y, there may be a dual representation which is invariant to Y but selective for X. In general, they

are both needed for different tasks and both can be computed by a HW module. In general, the machinery computing them shares a good deal of overlap. As an example, we would expect that different face patches in cortex are used to represent different combinations of invariance and selectivity.

2.1 Invariance reduces sample complexity of learning

Images of the same object usually differ from each other because of generic transformations such as translation, scale (distance) or more complex ones such as viewpoint (rotation in depth) or change in pose (of a body) or expression (of a face) (see also Anselmi et al. 2013, par 3.1.2 for a back of envelope estimation). In a machine learning context, invariance to image translations, for instance, can be built up trivially by memorizing examples of the specific object in different positions. Human vision on the other hand is clearly invariant for novel objects seen just once: people do not have any problem in recognizing in a distance-invariant way a face seen only once. It is rather intuitive that representations of images that are invariant to transformations such as scaling, illumination and pose, just to mention a few, should allow supervised learning from much fewer examples.

This conjecture is supported by previous theoretical work showing that almost all the complexity in recognition tasks is often due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the object (Lee and Soatto 2012). It implies that in many cases, recognition—i.e., both identification, e.g., of a specific car relative to other cars—as well as categorization, e.g., distinguishing between cars and airplanes—would be much easier (only a small number of training examples would be needed to achieve a given level of performance) if the images of objects were rectified with respect to all transformations, or equivalently, if the image representation itself were invariant. The case of identification is obvious since the difficulty in recognizing exactly the same object, e.g., an individual face, is only due to transformations. In the case of categorization, consider the suggestive evidence from the classification task in Figure 1.

The figure shows that if an oracle factors out all transformations in images of many different cars and airplanes, providing "rectified" images with respect to viewpoint, illumination, position and scale, the problem of categorizing cars vs airplanes becomes easy: it can be done accurately with very few labeled examples. In this case, good performance was obtained from a single training image of each class, using a simple classifier. In other words, the sample complexity of the problem seems to be very low.

A proof of the conjecture for the special case of translation or scale or rotation is provided in (Anselmi et al. 2014) for images defined on a grid of pixels the theorem (in the case of translations) can be stated as:

Sample complexity for translation invariance
Consider a space of images of dimensions pp which may appear in any position within a window of size $rp \times rp$. The natural image representation yields

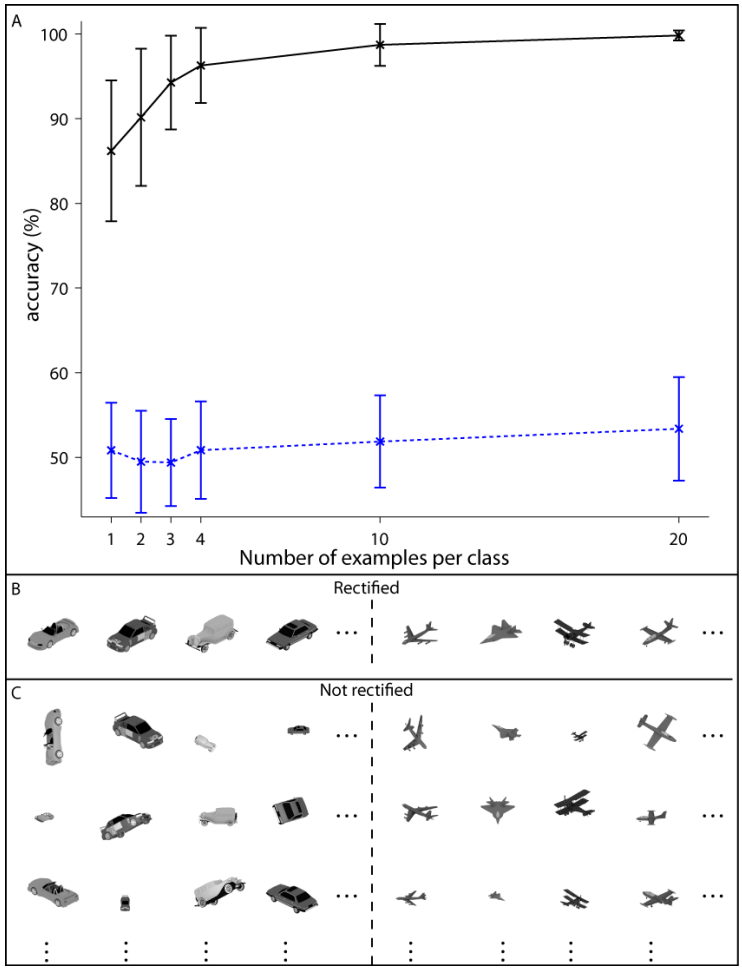

Figure 1: If an "oracle" factors out all transformations in images of many different cars and airplanes, providing "rectified" images with respect to viewpoint, illumination, position and scale, the problem of categorizing cars vs airplanes becomes easy: it can be done accurately with very few labeled examples. In the figure, good performance (black line) was obtained from a single training image from each rectified class, using a linear classifier operating on pixels, whereas training from the unrectified training set yields chance performance. In other words, the sample complexity of the problem becomes much lower with a rectified (and therefore invariant) representation (Anselmi et al. 2013).

a sample complexity (for a linear classifier) of order $m_{image} = O(r^2 p^2)$; the invariant representation yields a sample complexity of order:

$$m_{inv} = O(p^2)$$

The theorem says that an invariant representation can decrease considerably the sample complexity – that is, the number of supervised examples necessary for a certain level of accuracy in classification. A heuristic rule corresponding to the theorem is that the sample complexity gain is in the order of the number of virtual examples generated by the action of the (locally compact) group on a single image (see also Niyogi 1998, Y. S. Abu-Mostafa 1993). This is not a constructive result but it supports the hypothesis that the ventral stream in visual cortex tries to approximate such an oracle. The next section describes a biologically plausible algorithm that the ventral stream may use.

2.2 Unsupervised learning and computation of an invariant signature (one layer architecture)

The following HW algorithm is biologically plausible - as we will discuss in detail in section 6, where we argue that it may be implemented in cortex by a HW module, that is a set of KH complex cells with the same receptive field, each pooling the output of a set of simple cells whose sets of synaptic weights correspond to one of the K "templates" of the algorithm and its transformations (which are also called templates) and whose output is filtered by a sigmoid function with Δh threshold, $h = 1, , H$.

HW algorithm for (locally compact) groups (see Figure 2)

- "Developmental" stage:
  1.1 For each of K isolated (on an empty background) objects - "templates" - memorize a sequence Γ of $|G|$ frames corresponding to its transformations $(g_i, i = 1, ..., |G|)$ observed over a time interval (thus $\Gamma = g_0 t, g_1 t ..., g_{|G|} t$ for template t; for template t^k the corresponding sequence of transformations is denoted Γ_k).
  1.2 Repeat for each of K templates

- "Run-time" computation of invariant signature for a single image (of any new object):
  2.1 For each Γ_k compute the dot product of the image with each of the $|G|$ transformations in Γ_k
  2.2 For each k compute cumulative histogram of the resulting values
  2.3 The signature is the set of K cumulative histograms that is the set of:

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + h\Delta) \tag{1}$$

  where I is an image, σ is a threshold function $\Delta > 0$ is the width of bin in the histogram and $h = 1, , H$ is the index of the bins of the histogram.

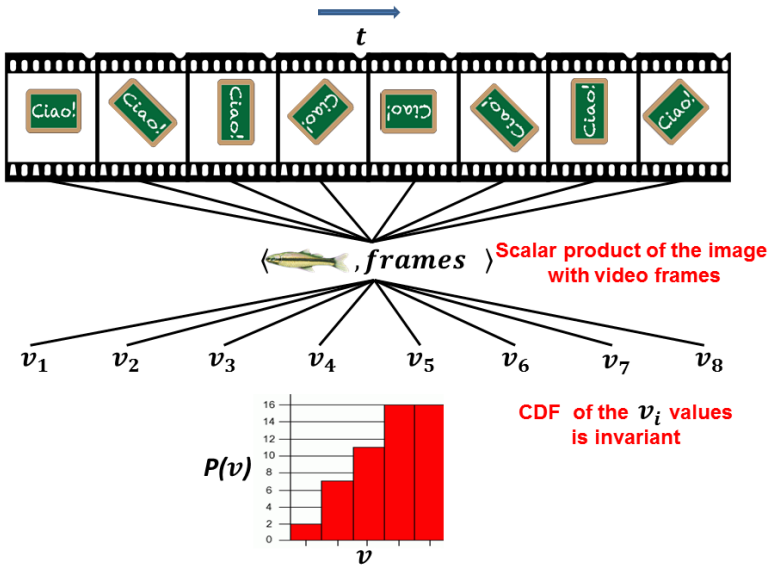

Figure 2: A graphical summary of the HW algorithm. The set of $\mu_h^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + h\Delta)$ values (eq. 1) in the main text) correspond to the the histogram where k=1 denotes the template" green blackboard", h the bins of the histogram and the transformations are from the rotation group. Crucially, mechanisms capable of computing invariant representations under affine transformations can be learned (and maintained) in an unsupervised, automatic way just by storing sets of transformed templates which are unrelated to the object to be represented in an invariant way. In particular the templates could be random patterns..

The algorithm consists of two parts: the first is unsupervised learning of transformations by storing transformed templates, which are "images". This can be thought of as a "once in a time" stage, possibly done once during development of the visual system. The second part is the actual computation of invariant signatures during visual perception.

This is the algorithm used throughout the paper. The guarantees we can provide depend on the type of transformations. The main questions are a) whether the signature is invariant under the same type of transformations that were observed in the first stage and b) whether it is selective, e.g. it can distinguish between N different objects. The summary of the main results of (Anselmi et al. 2014, Anselmi et al. 2013) is that the HW algorithm is invariant and selective (for K in the order of $\log N$) if the transformations form a group. In this case, any set of randomly chosen templates will work for the first stage. Seen as transformations from a 2D image to a 2D image, the natural choice is the affine group consisting of translations, rotations in the image plane, scaling (possibly non-isotropic) and compositions thereof of (see also par 3.2.3 of Anselmi et al 2013). The HW algorithm can learn with exact invariance and desired selectivity [1] in the case of the affine group or its subgroups. In the case of 3D "images" consisting of voxels with x, y, z coordinates, rotations in 3D are also a group and in principle can be dealt with, achieving exact invariance from generic templates by the HW algorithm (in practice this is rarely possible because of correspondence problems and self-occlusions). Later in section 2.5 we will show that the same HW algorithm provides approximate invariance (under some conditions) for non-group transformations such as the transformations from R^3 to R^2 induced by 3D rotations of an object.
In the case of compact groups the guarantees of invariance and selectivity are provided by the following two theorems (given informally here; detailed formulation in Anselmi et al. 2014, Anselmi et al. 2013).

Invariance theorem
The distributions represented by equation 1 are invariant, that is each bin is invariant, e.g.

$$\mu_h^k(I) = \mu_h^k(gI) \tag{2}$$

for any g in G, where G is the (locally compact) group of transformations labeled g_i in equation 1.

Selectivity theorem
For (locally compact) groups of transformations (such as the affine group), the

[1] Consider the case of a discrete group of M elements. What is required are K templates with $K = M + 1$. The argument is as follows. From the M observations there is a distribution $p(I)$ supported on M atoms (each is a bin corresponding to a specific image). This distribution can be uniquely associated to $M + 1$ one-dimensional distributions of the projections of $p(I)$. This is based on Heppes theorem (Heppes 1956): A distribution consisting of k arbitrary points in the n-dimensional space is uniquely determined if its projections on $k + 1$ not parallel 1-dimensional subspaces are given.

distributions represented by equations 1) can achieve any desired selectivity for an image among N images in the sense that they can ϵ-approximate the true distance between each pair of the images (and any transform of them) with probability $1 - \delta$ provided that

$$K > \frac{c}{\epsilon^2} \ln \frac{N}{\delta} \tag{3}$$

where c is a universal constant.
The signature provided by the K cumulative histograms is a feature vector corresponding to the activity of the (HK) complex cells associated with the HW module. It is selective in the sense that it corresponds uniquely to an image of a specific object independently from its transformation. It should be noted that the robustness or stability of the signature under noisy measurements remains an interesting open problem in the theory. Because of the restricted dynamic range of cortical cells the number H of bins is likely to be small, probably around 2 or 3. It is important to remark that other, related representations are possible (see also par. 3.3.1 eq. 3 Anselmi et al. 2013 and Kouh and Poggio 2008). A cumulative distribution function (cdf) is fully represented by all its moments; often a few moments, such as the average or the variance (energy model of complex cells, see Adelson and Bergen 1985) or the max,

$$\begin{aligned} \mu^k_{av}(I) &= \frac{1}{|G|} \sum_{i=1}^{|G|} \left\langle I, g_i t^k \right\rangle \\ \mu^k_{energy}(I) &= \frac{1}{|G|} \sum_{i=1}^{|G|} \left\langle I, g_i t^k \right\rangle^2 \\ \mu^k_{max}(I) &= max_{g_i \in G} \left\langle I, g_i t^k \right\rangle \end{aligned} \tag{4}$$

can effectively replace the cumulative distribution function. Notice that any linear combination of the moments is also invariant and a small number of linear combinations is likely to be sufficiently selective. We will discuss implications of this remark for models of complex cells in the last section.

2.3 Optimal templates for scale and position invariance are Gabor functions

The previous results apply to all groups, in particular to those which are not compact but only locally compact such as translation and scaling. In this case it can be proved that invariance holds within an observable window of transformations (Anselmi et al. 2014, Anselmi et al. 2013). For the HW module the observable window corresponds to the receptive field of the complex cell (in space and scale). For maximum range of invariance within the observable window, it is proven in (Anselmi et al. 2014, Anselmi et al. 2013) that the templates must be maximally sparse relative to generic input images (see below for definition of sparseness). In the case of translation and scale invariance, this sparsity requirement is equivalent to localization is space and spatial frequency,

respectively: templates must be maximally localized for maximum range of invariance - in order to minimize boundary effects due to the finite window. Assuming therefore that the templates are required to have simultaneously a minimum size in space and spatial frequency, it follows from results of Gabor (Gabor 1946, see also Donoho 1989) that they must be Gabor functions. The following surprising property holds:

Optimal invariance theorem Gabor functions of the form (here in 1D) $t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}$ are the templates that are simultaneously maximally invariant for translation and scale (at each x and ω.)
In general, templates chosen at random in the universe of images can provide scale and position invariance. However, for optimal invariance under scaling and translation, templates of the Gabor form are optimal. This is the only computational justification we know [2] of the Gabor shape of simple cells in V1 which seems to be remarkably universal: it holds in primates (Ringach 2002), cats (Jones et al. 1987) and mice (Striker and Neill 2008) (see also Figure 3 for results of simulations).

2.4 Quasi-invariance to non-group transformations requires class-specific templates

All the results so far require a group structure and provide exact invariance for a single new image. In 2D this induces all combination of translation, scaling and rotation in the image plane but does not include the transformations induced on the image plane by 3D transformations such as viewpoint changes and rotation in depth of an object. The latter forms a group in 3D, that is if images and templates were 3D views: in principle motion or stereopsis can provide the third dimension though available psychophysical evidence (Bulthoff and Edelman 1992, Tarr 1995) suggests that human vision does not use it for recognition. Notice that transformations in the image plane are affected not only by orthographic projection of the 3D geometry but also by the process of image formation which depends on the 3D geometry of the object, its reflectance properties, the relative location of light source and viewer.

It turns out that the HW algorithm can still be applied to non-group transformations – such as transformations of the expression of a face, of pose of a body–to provide, under certain conditions, approximate invariance around the "center" of such a transformation. In this case bounds on the invariance depend on specific details of the object and the transformation: we do not have general results and suspect they may not exist. The key technical requirement is that a new type of sparsity condition holds: sparsity for the class of images I_C with respect to the dictionary t^k under the transformations T_r (we consider

[2] Mallat (Mallat 2012) justification of his use of wavelets is different (Lipschitz-continuity to the action of C^2 diffeomorphisms). He does not justify the Gaussian envelope associated to Gabor functions. Also the alternative argument of Stevens 2004 for wavelet receptive fields in V1 does not imply Gabor wavelets.

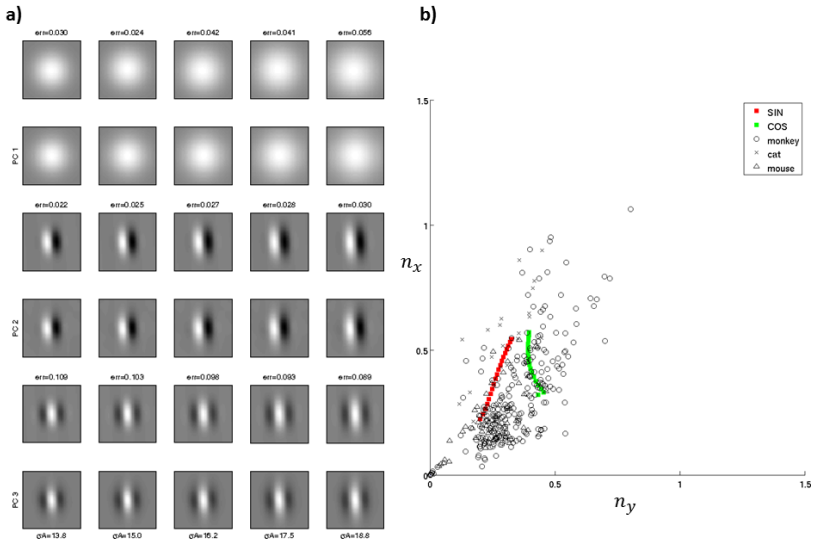

Figure 3: a) Simulation results for V1 simple cells learning via PCA. Each "cell" sees a set of images through a Gaussian window (its dendritic tree) ., shown in the top row. Each cell then "learns" the same weight vector extracting the principal components of its input. b) This figure shows $n_y = \sigma_y/\lambda$ vs $n_x = \sigma_x/\lambda$ for the modulated (x) and unmodulated (y) direction of the Gabor wavelet. Notice that the slope σ_y/σ_x Is a robust finding in the theory and apparently also in the physiology data. Neurophysiology data from monkeys, cats and mice are reported together with our simulations. Figure from Poggio et al. 2013.

here a one parameter r transformation)

$$\left\langle I_C, T_r t^k \right\rangle \approx 0 \quad |r| > a \quad a \gtrsim 0. \tag{5}$$

This property, which is an extension of the compressive sensing notion of "incoherence", requires that images in the class and the templates have a representation with sharply peaked correlation and autocorrelation (the constant a above is related to the support of the peak of the correlation). This condition can be satisfied by templates that are similar to images in the set and are sufficiently "rich" to be incoherent for "small" transformations. This relative sparsity condition is usually satisfied by the neural representation of images and templates at some high level of the hierarchy of HW modules that we describe next. Like standard sparsity (Donoho, 1989) our new sparsity condition is generic: most neural patterns - templates and images from the same class - chosen at random will satisfy it. The full theorem (Anselmi et al. 2014, Anselmi et al. 2013) takes the following form:

Class-specific property
$\mu_h^k(I)$ is approximatively invariant around a view if

- I is sparse in the dictionary of the templates relative to the transformations
- I transforms "in the same way" as the templates
- the transformation is smooth.

The main implication is that approximate invariance can be obtained for non-group transformation by using templates specific to the class of objects. This means that class specific modules are needed, one for each class; each module requires highly specific templates, that is tuning of the cells. The obvious example is face-tuned cells in the face patches. Unlike exact invariance for affine transformations where tuning of the "simple cells" is non-specific in the sense that does not depend on the type of image, non-group transformations require highly tuned neurons and yield at best only approximate invariance [3]

[3] A similar, stronger result can be obtained with somewhat more specific assumptions. Consider the image transformations induced by an affine transformation of a 3D object such as a face, in particular rotation around the vertical (y) axis. Consider a simplistic model of image formation in which the surface texture of the 3D object is mapped into pixel values independently of the surface orientation and of the geometry of illumination. Let us call $I^0(x)$, with $x = (x, y)$ the image of the object for "zero" pose. Let the object rotated by θ yield the image $I^\theta(x)$. Introducing the deformation field $d^\theta(x)$ we write its definition as $I^\theta(x) = I^0(x + d^\theta(x))$. The deformation field $d(x)$ is uniquely associated with the 3D structure of a specific object (under orthographic projection and neglecting occlusions, the field $d^\theta(x) = \sum_{i=1}^{2} c_i^\theta \phi_i(x)$, where $\phi_1(x)$ and $\phi_2(x)$ are fixed functions for a specific object, describes the deformation on the image plane induced by affine transformations in 3D (this follows since under orthographic projection the space of views of the object is spanned by 2 views, see Ull-

2.5 Two stages in the computation of an invariant signature

Hierarchical architectures are advantageous for several reasons which are formalized mathematically in (Anselmi et al. 2014, Anselmi et al. 2013). It is illuminating to consider two extreme "cartoon" architectures for the first of the two stages described at the beginning of section 2:

- one layer comprising one HW module and its KH complex cells, each one with a receptive field covering the whole visual field
- a hierarchy comprising several layers of HW modules with receptive fields of increasing size, followed by parallel modules, each devoted to invariances for a specific object class.

In the first architecture invariance to affine transformations is obtained by pooling over KH templates each one transformed in all possible ways: each of the associated simple cells corresponds to a transformation of a template. Invariance over affine transformation is obtained by pooling over the whole visual field. In this case, it is not obvious how to incorporate invariance to non-group transformations directly in this one-hidden layer architecture.

Notice however that a HW module dealing with non-group transformations can be added on top of the affine module. The theorems (Anselmi et al. 2014, Anselmi et al. 2013) allow for this factorization. Interestingly they do not allow in general for factorization of translation and scaling (e.g. one layer computing translation invariance and the next computing scale invariance). Instead, what the mathematics allows is factorization of the range of invariance for the same group of transformations (see also Anselmi et al. 2013 par 3.6-7-8-9). This justifies the first layers of the second architecture, corresponding to Figure 4 stage1, where the size of the receptive field of each HW module and the range of its invariance increases from lower to higher layers.

---

man and Basri 1991; Poggio 1990.). Under our (strong) assumptions, the normalized images $I_1^\theta(x), I_2^\theta(x)$ of two objects with different texture but the same 3D structure - therefore the same deformation fields – satisfy $\langle I_1^0(x), I_2^0(x)\rangle = \langle I_1^\theta(x), I_2^\theta(x)\rangle$. Objects with similar texture and similar 3D structure and therefore similar deformation field therefore satisfy $\langle I_1^0(x), I_2^0(x)\rangle \approx \langle I_1^\theta(x), I_2^\theta(x)\rangle$. This property of objects that belong to the same object class (Anselmi et al. 2014, Anselmi et al. 2013 and Leibo et al. 2014) is sufficient for the signature component $\mu_h^k(I) = (1/|G|)\sum_\theta \sigma(< I, t^\theta > +h\Delta)$ to be approximately invariant under 3D rotation by θ. The property holds for higher layers in a hierarchical architecture under mild conditions (Anselmi et al in preparation). This in turn can explain how templates from the object class of faces can be used to obtain invariance to rotations in depth of a new face under assumptions similar to the conditions required by the class-specific theorem (apart from sparsity, which is not strictly required).

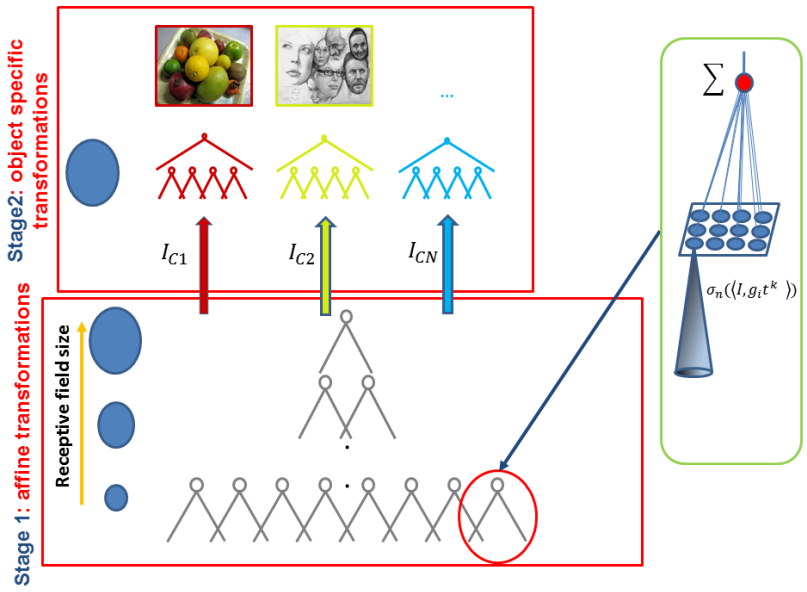

Figure 4: A hierarchical architecture of HW modules. Signature provided by each of the nodes at each layer may be used by a supervised classifier. Stage 1: a hierarchy of HW modules (green inset) with growing receptive fields provide a final signature (top of the hierarchy) which is globally invariant to affine transformations by pooling over a cascade of locally invariant signatures at each layer. Stage 2: transformation specific modules provide invariance for non group transformations (e.g. rotation in depth).

2.6 Invariance to translation and scale (stage 1) tolerant of clutter requires a hierarchical architecture

The main problem with the one-layer architecture is that it can recognize isolated objects in the visual field in an invariant way but cannot recognize objects in clutter: *the key theorem about invariance assumes that image and templates portray isolated objects*. Otherwise the signature may change because of different clutter at different times[4] . The problem of clutter - of recognizing an object independently of the presence of another one nearby - is closely related to the problem of recognizing "wholes" and "parts". Recognizing an eye in a face has the problem that the rest of the face is clutter. This is the old conundrum of recognizing a tree in a forest while still recognizing the forest.

A partial solution to this problem is a hierarchical architecture for stage 1 in which lower layers provide signatures with a small range of invariance for "small" parts of the image and higher layers provide signatures with greater invariance for larger parts of the image (see also Anselmi et al. 2013, par 5.4 for the case of translations). This signature could then be used by class specific modules. We will describe this architecture starting with the retina and V1 in the next section. Two points are of interest here.

Factorization of range of invariances is possible if a certain property of the hierarchical architecture, called covariance, holds. Assume a group transformation of the image that is e.g. a translation or scaling of it. The first layer in a hierarchical architecture is called covariant if the pattern of neural activity at the output of the complex cells transforms accordingly to the same group of transformations. It turns out that the architectures we describe have this property (see Anselmi et al. 2014, Anselmi et al. 2013 par 3.5.3 for the translations case): isotropic architectures, like the ones considered in this paper, with point-wise nonlinearities, are covariant. Since each module in the architecture gives an invariant output if the transformed object is contained in the pooling range, and since the pooling range increase from a layer to the next, there is a invariance over larger and larger transformations. The second point is that in order to make recognition possible for both parts and wholes of an image, the supervised classifier should receive signatures not only from the top layer (as in most neural architectures these days) but from the other levels as well (directly or indirectly).

[4]Notice that because images are filtered by the retina with spatial bandpass filters (ganglion cells), the input to visual cortex is a rather sparse pattern of activities, somewhat similar to a sparse edge map.

3 Biophysical mechanisms of invariance: unsupervised learning, tuning and pooling

3.1 A single cell model of simple and complex cells

There are at least two possible biophysical models for the HW module implied by i-theory. The first is the original Hubel and Wiesel model of simple cells feeding into a complex cell. I-theory proposes the "ideal" computation of a CDF, in which case the nonlinearity at the output of the simple cells is a threshold. A complex cell, summating the outputs of a set of simple cells, would then represent a bin of the histogram; a different complex cell in the same position pooling a set of similar simple cells with a different threshold would represent another bin of the histogram. Another possibility, is that the nonlinearity at the output of the simple cells is a square or any power or combination of powers. In this case the complex cell pooling simple cells with the same nonlinearity would represent a moment of the distribution, including the linear average. Notice that in this case some of the complex cells would be linear and would be classified by neurophysiologists using the standard criteria as simple! The nonlinear transformation at the output of the simple cells would correspond to the spiking mechanism in populations of cells (see references in Kouh and Poggio 2008).

The second biophysical model for the HW module that implements the computation required by i-theory consists of a single cell where dendritic branches play the role of simple cells (each branch containing a set of synapses with weights providing, for instance, Gabor-like tuning of the dendritic branch) with inputs from the LGN; active properties of the dendritic membrane distal to the soma provide separate threshold-like nonlinearities for each branch separately, while the soma summates the contributions for all the branches. This model would solve the puzzle that so far there seems to be no morphological difference between pyramidal cells classified as simple vs complex by physiologists.

It is interesting that i-theory is robust with respect to the nonlinearity from simple to the complex "cells". We conjecture that almost any set of non trivial nonlinearities will work. The argument rests on the fact that a set of different complex cells pooling from the same simple cells should compute the cumulative distribution or equivalently its moments or combinations of moments (each combination is a specific nonlinearity). Any nonlinearity will provide invariance, if the nonlinearity does not change with time and is the same for all the simple cells pooled by the same complex cells. A sufficient number of different nonlinearities, each corresponding to a complex cell, can provide appropriate selectivity - assuming that each nonlinearity can be represented by a truncated power series and that the associated complex cells provide therefore enough linearly independent combinations of moments.

3.2 Learning the wiring

A simple possibility of how the wiring between a group of simple cells with the same tuning (for instance representing the same eigenvector, with the same orientation etc.) and a complex cell may develop is to invoke a Hebbian trace rule (Foldiak 1991). In a first phase complex cells may have subunits with different selectivities (eg orientations), for instance because natural images are rotation invariant and thus eigenvectors with different orientations are degenerate. In a second plastic phase, subunits which are not active when the majority of the subunit is active will be pruned out according to a Foldiak-like rule.

3.3 Hebb synapses and PCAs

I-theory provides the following algorithm for learning the relevant invariances during unsupervised visual experience: storing sequences of images for each of a few objects (called "templates") while transforming - for instance translating, rotating and looming. Section 2 proves that in this way invariant hierarchical architectures can be learned from unsupervised visual experience. Such architectures represent a significant extension, beyond simple translation invariance and beyond hardwired connectivity, of models of the ventral stream such as Fukushima's Neocognitron (Fukushima1980) and HMAX (Riesenhuber and Poggio 2000, Serre et al. 2007) – as well as deep neural network called convolutional networks (LeCun et al. 1989, LeCun et al. 1995) and related models e.g. (Poggio et al. 1990, Perrett and Oram 1993, Mel 1997, Stringer and Rolls 2002, Pinto et al. 2009, Saxe et al. 2011, Le et al. 2011, Abel-Hamid 2012).

In biological terms the sequence of transformations of one template would correspond to a set of simple cells, each one storing in its tuning a frame of the sequence. In a second learning step a complex cell would be wired to those "simple" cells. However, the idea of a direct storage of sequences of images or image patches in the tuning of a set of V1 cells by exposure to a single object transformation is biologically rather implausible. Since Hebbian-like synapses are known to exist in visual cortex a more natural hypothesis is that synapses would incrementally change over time as an effect of the visual inputs - that is over many sequences of images resulting from transformations of objects, e.g. templates. The question is whether such a mechanism is compatible with i-theory and how.

We explore this question for V1 in a simplified setup that can be extended to other areas. We assume

- a) that the synapses between LGN inputs and (immature) simple cells are Hebbian and in particular that their dynamics follows Oja's flow. In this case, the synaptic weights will converge to the eigenvector with the largest eigenvalue of the covariance of the input images.

- b) that the position and size of the untuned simple cells is set during development according to the inverted pyramidal lattice of Figure 7. The

key point here is that the size of the Gaussian spread of the synaptic inputs and the positions of the ensemble of simple cells are assumed to be set independently of visual experience.

In summary we assume that the neural equivalent of the memorization of frames (of transforming objects) is performed online via Hebbian synapses that change as an effect of visual experience. Specifically, we assume that the distribution of signals "seen" by a maturing simple cell is Gaussian in x, y reflecting the distribution on the dendritic tree of synapses from the lateral geniculate nucleus. We also assume that there is a range of Gaussian distributions with different σ which increase with retinal eccentricity. As an effect of visual experience the weights of the synapses are modified by a Hebb rule (Hebb 1949). Hebb's original rule can be written as

$$w_n = \alpha y(x_n) x_n \tag{6}$$

where α is the "learning rate", x_n is the input vector w is the presynaptic weights vector and y is the postsynaptic response. In order for this dynamical system to actually converge, the weights have to be normalized. In fact, there is considerable experimental evidence that cortex employs normalization (Turrigiano and Nelson 2004) and references therein). Hebb's rule appropriately modified with a normalization factor turns out to be an online algorithm to compute PCA from a set of input vectors. In this case it is called Oja's flow. Oja's rule (Oja 1982, Karhunen 1994) defines the change in presynaptic weights w given the output response y of a neuron to its inputs to be

$$\Delta w_n = w_{n+1} - w_n = \alpha y_n (x_n - y_n w_n) \tag{7}$$

where $y_n = w_n^T x_n$. The equation follows from expanding to the first order Hebb rule normalized to avoid divergence of the weights.
Since the Oja flow converges to the eigenvector of the covariance matrix of the x_n which has the largest eigenvalue, we are therefore led to analyze the spectral properties of the inputs to "simple" cells and study whether the computation of PCA can be used by the HW algorithm and in particular whether it satisfies the selectivity and invariance theorems.

Alternatives to the Oja's rule that still converge to PCAs can be considered (Sanger 1989 and Oja 1992). Also notice that a relatively small change in the Oja equation gives an online algorithm for computing ICAs instead of PCAs (see Hyvrinen and Oja 2000). Which kind of plasticity is closer to the biology remains an open question.

3.4 Spectral theory and pooling

Consider stage 1, which is retinotopic, and, in particular, the case of simple cells in V1. From assumption b in section 6.1, the lattice in x, y, s of immature simple cell is set during development of the organism (s is the size of the Gaussian envelope of the immature cell). Assume that all of the simple cells are exposed,

while in a plastic state, to a possibly large set of images $T = (t_1, ..., t_K)$. A specific cell at a certain position in x, y, s is exposed to the set of transformed templates g_*T where g_* corresponds to the translation and scale that transforms the "zero" cells to the chosen neuron in the lattice) and therefore the associated covariance matrix $g_*TT^Tg_*^T$. Thus it is possible to choose PCA as new templates and pooling over corresponding PCAs across different cells is equivalent to pool over a template and its transformations. Both the invariance and selectivity theorem are valid. Empirically, we find (Leibo et al. 2014) that PCA of natural images provides eigenvector that are Gabor-like wavelets with a random orientation for each size receptive field (hypothesis b). The random orientation is because of the argument above, together with the fact that the covariance of natural images is approximately rotation invariant. The Gabor-like shape can be qualitatively explained in terms of translation invariance of the correlation matrix associated with a set or natural images (and their approximate scale invariance which corresponds to a $\approx 1/f$ spectrum, see also Ruderman 1994 and Torralba and Oliva 2003)[5] . Thus the Oja rule acting on natural images provides "equivalent templates" that are Gabor-like - which are the optimal ones, according to the theory of section 2.3!

Consider now non-retinotopic stage 2 in which transformations are not in scale or position, such as the transformation induced by a rotation of a face. Assume that a "simple" cell is exposed to "all" transformations g_i (g_i is a group element of the finite group G) of each of a set $T = (t_1, ..., t_K)$ of K templates. The cell is thus exposed to a set of images (columns of X) $X = (g_1T, ..., g_{|G|}T)$. For the sake of this example, assume that G is the discrete equivalent of a group. Then the covariance matrix determining the Oja's flow is

$$C = XX^T = \sum_{i=1}^{|G|} g_i TT^T g_i^T. \tag{8}$$

It is immediate to see that if ϕ is an eigenvector of C then $g_i\phi$ is also an eigenvector with the same eigenvalue (for more details on how receptive fields look like in V1 and higher layers see also Poggio et al. 2013 or Anselmi et al. par 4.3.1 and 4.7.3 or Gallant et al. 1993,1996 or Hegde and Van Essen 2000). Consider for example G to be the discrete rotation group in the plane: then all the (discrete) rotations of an eigenvector are also eigenvectors. The Oja rule will converge to the eigenvectors with the top eigenvalue and thus to the subspace spanned by them. It can be shown that L^2 pooling over the PCA with the same eigenvalues represented by different simple cells is then equivalent to L^2 pooling over transformations, as the theory of section 2.2 dictates in order to

[5]Suppose that the simple cells are exposed to patterns and their scaled and translated versions. Suppose further that images are defined on a lattice and translations and scaling (a discrete similitude group) are carefully defined on the same lattice. Then a set of discrete orthogonal wavelets - defined in terms of discrete dilation and shifts - exist and is invariant under the group. The Oja rule (extended beyond the top eigenvector) could converge to specific wavelets.

achieve selectivity and invariance (Anselmi et al. 2013 par 4.6.1). This argument can be formalized in the following variation of the pooling step in the HW algorithm:

Spectral pooling proposition. Suppose that ϕ_k is the matrix corresponding to the group transformations of template t^k (each column is a transformation of the template). Consider the set of eigenvectors ϕ_i^* of the associated covariance matrix with eigenvalue λ^*. Because of the above argument $\langle g_m I, \phi_k^* \rangle = \langle I, \phi_p^* \rangle$ where $g_m^{-1}\phi_k^* = \phi_p^*$. Therefore to achieve invariance a complex cell can pool with a quadratic nonlinearity over the eigenvectors of Γ_k instead than over the transformations of the template (in addition we may assume sparsity of the eigenvectors). The argument is still valid if the pooling is over part of the eigenvectors of $\Gamma = \bigcup_{i=1}^{K} \Gamma_i$. Thus components of an invariant signature can be computed as

$$\mu_*(I) = \sum_i || \langle I, \phi_i^* \rangle ||^2. \tag{9}$$

We conjecture that PCA and linear combinations of them can provide a sufficient number of templates to satisfy the selectivity theorem of section 2.

3.5 Tuning of "simple" cells

The theorems of section 2 on the HW module, imply that the templates, and therefore the tuning of the simple cells can be the image of any object. At higher levels in the hierarchy, the templates are neuroimages - patterns of neural activity – induced by actual images in the visual field. The previous section, however, offers a more biologically plausible way to learn the templates from unsupervised visual experience, via Hebbian plasticity. In the next sections we will discuss predictions following from this assumptions for the tuning of neurons in the various areas of the ventral stream.

4 Stage 1: retina and V1

4.1 Inverted truncated pyramid

The simplest and most common image transformations are similitude transformations, that is shifts in position (x) and uniform changes in scale (s). The theory suggests that the first step of the invariance computation are likely to consists of learning/storage of the set of transformations. This is done either by actual visual experience or by evolution encoded in the genes or by a combination (see the Discussion section):

- storing template t^k (which is an image patch and could be chosen at random)

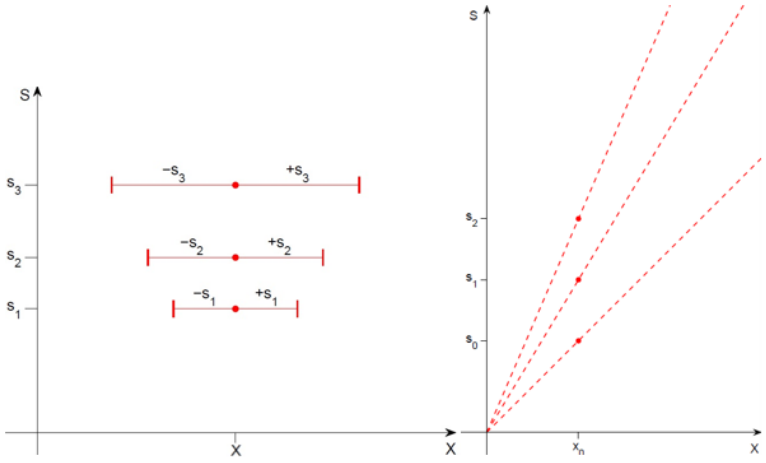

Figure 5: On the left the s, x plane, indicating templates of radius s_1, s_2, and s_3 in the spatial dimensions x, y all centered at position X. Note that while s and x are both measured in degrees of visual angle, in this plot the two axes are not shown to the same scale. Here, as in the rest of the paper, we show only one spatial coordinate (x); everything we say can be directly extended to the x, y plane. We will assume later that the smallest template is the smallest simple cell in the fovea with a radius of around $40''$ (Marr et al. 1980). For any fixed eccentricity (right) the size of the pattern determines the slope of its s,x trajectory under scaling. (From Poggio et al., 2014).

- storing all its observed transformations (bound together by continuity in time)
- repeating the above process for a set of K templates.

Though the templates can be arbitrary image patches, we will assume – because of the Optimal Invariance Theorem and because of the experimental evidence from V1 simple cells – that V1 templates are Gabor-like functions (more precisely windowed Fourier transforms, Mallat 2008, par 4.2 pg. 92).

Under scaling, a pattern exactly centered at the center of the fovea will change size without any translation of its center while its boundaries will shift in x, y; for a pattern centered at some non-zero eccentricity, scaling will translate its center in the s, x plane, see Figures 5. In the s, x plane the slope of the trajectory of a pattern under scaling is a straight line through the origin with a slope that depends on the size of the pattern s and the associated position.

Consider a Gabor function at $s = s_0, x = 0$. s_0 is the minimum possible receptive field size given optical constraints. Transforming it by shifts in x within $(-x_0, x_0)$ generates a set Γ_0 of templates. Suppose that we want to ensure that what is recognizable at the highest resolution (s_0) remains recognizable at all

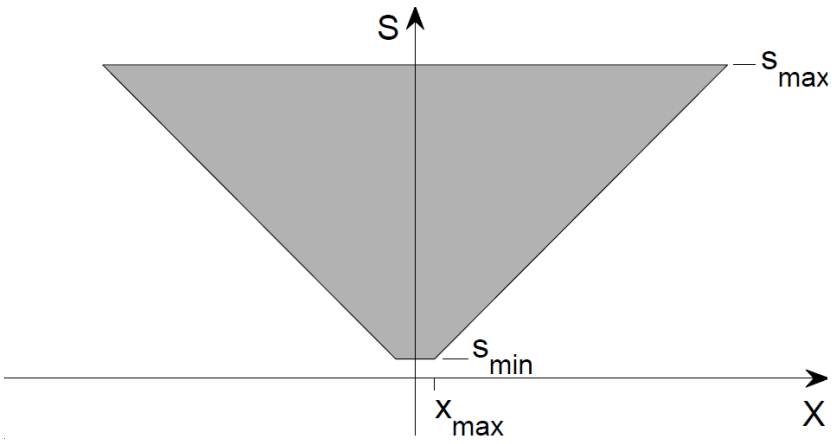

Figure 6: Synthesis of template set. Consider a template, such as a Gabor RF, at s_{min} and $x = 0$. Store its transformations under (bounded) shift, filling the interval between s_{min}, $x = 0$ and s_{min} ,x_{max}. Store its transformations over (bounded) scale, filling the space in the truncated, inverted pyramid shown in the figure. This is the space of bounded joint transformations in scale and space. For clarity of the figure the axes s,x are not in the same units (s unit is 10 times the x unity, so that the real slope is $1/10$ of the shown one). (From Poggio et al., 2014).

scales up to s_{max}. The associated scale transformations of the set Γ_0 yield the inverted truncated pyramid shown in Figure 6.

Pooling over that set of transformed templates according to the HW algorithm will give uniform invariance to all scale transformations of a pattern over the range (s_0, s_{max}); invariance to shifts will be at least within $(-x_0, x_0)$, depending on scale. Note that the above process of observing and storing a set of transformations of templates in order to be able to compute invariant representations may take place at the level of evolution or at the level of development of an individual system or as a combination of both.

The following definition holds: the inverted truncated pyramid of Figure 6 is the locus Γ of the points such that their scaling between s_{min} and s_{max} gives points in S; further all points P in $(-x_0, x_0)$ are in S, e.g. $P \in \Gamma$ if

$$g_s P \in \Gamma \quad , s \in (s_{min}, s_{max}), \Gamma_0 \in \Gamma \tag{10}$$

where Γ_0 consists of all points at s_{min} between $-x_0$ and x_0. (Other alternatives are possible: one of many is to set a constant difference between the minimum and the maximum scale at each eccentricity: $s_{max} - s_{min} = const$ (see Figure 9 (lower) of Poggio et al., 2014). Experimental data suggests this is a more likely possibility (large eccentricities are also represented in the visual system)).
The inverted pyramid region follows naturally if scale invariance has a higher priority than shift invariance. We recall that according to the Optimal invari-

ance theorem, the optimal template is a Gabor function (Anselmi et al. 2014, Anselmi et al. 2013). Under scale and translation within the inverted pyramid region the Gabor template originates a set of Gabor wavelets (a tight frame). In the region of Figure 6 scaling each wavelet generates other elements of the frame. Notice that the shape of the lower boundary of the inverted pyramid, though similar to standard empirical functions that have been used to describe M scaling, is different for small eccentricities. An example of an empirical function (Cowey et al., 1974), described in (Strasburger et al., 2011) is $M^{-1} = M^0(1 + ax)$, where M is the cortical magnification factor, M_0 and a are constants and x is eccentricity.

A normal, non-filtered image may activate all or part of the Gabor filters within the inverted truncated pyramid of Figure 6. The pattern of activities is related to a multi-resolution wavelet decomposition of an image. We call this transform of the image an"inverted pyramid scale-space fragment" ("IP fragment" in short: the term fragment is borrowed from Ullman) and consider it as supported on a domain in the space x, y, s that is contained in the inverted truncated pyramid of the figure. The fragment corresponding to a bandpass filtered image should be a more or less narrow horizontal slice in the s, x plane. In the following we assume that the template is a Gabor filter (of one orientation; other templates may have different orientations). We assume that the Gabor filter and its transforms under translation and scaling are roughly bandpass and *the sampling interval at one scale over* x is s implying half overlap of the filters in x. This is illustrated in Figure 7. These assumptions imply that for each array of filters of size s, the first unit on the right of the central one is at $x = s$, if x and s are measured in the same units. For the scale axis we follow the sampling pattern estimated by Marr et al., 1980 with five "frequency channels" with $2s = 1'20'', 3.1', 6.2', 11.7', 21'$. Filter channels as described above are supported by sampling by photoreceptors that starts in the center of the fovea at the Shannon rate, dictated by the diffraction-limited optics with a cutoff around 60 cycles/degree, and then decreases as a function of eccentricity.

4.2 Fovea and foveola

In this "truncated pyramid" model of the simple cells in V1, the slope of the magnification factor M as a function of eccentricity depends on the size of the foveola – that is the region at the minimum scale s_{min}. The larger the foveola, the smaller the slope. We submit that this model nontrivially fits data about the size of the fovea, the slope of M and other data about the size of receptive fields in V1. In particular the size of the foveola, the size of the largest RFs in AIT and the slope of acuity as a function of eccentricity depend on each other: fixing one determines the other (after setting the range of spatial frequency channels, i.e., the range of RF sizes at $x = 0$ in V1). For a back of the envelope calculation we assume here that $s_{min} \approx 40$" (from an estimate of $1'20''$ for the diameter of the smallest simple cells Marr et al. 1980, see also Mazer et al. 2002). Data of Hubel and Wiesel (Hubel et al., 1962, Hubel et al., 1974) (shown in figure 6A in Hubel and Wiesel 1974) and Gattass (Gattass et al. 1981, Gattass et al.

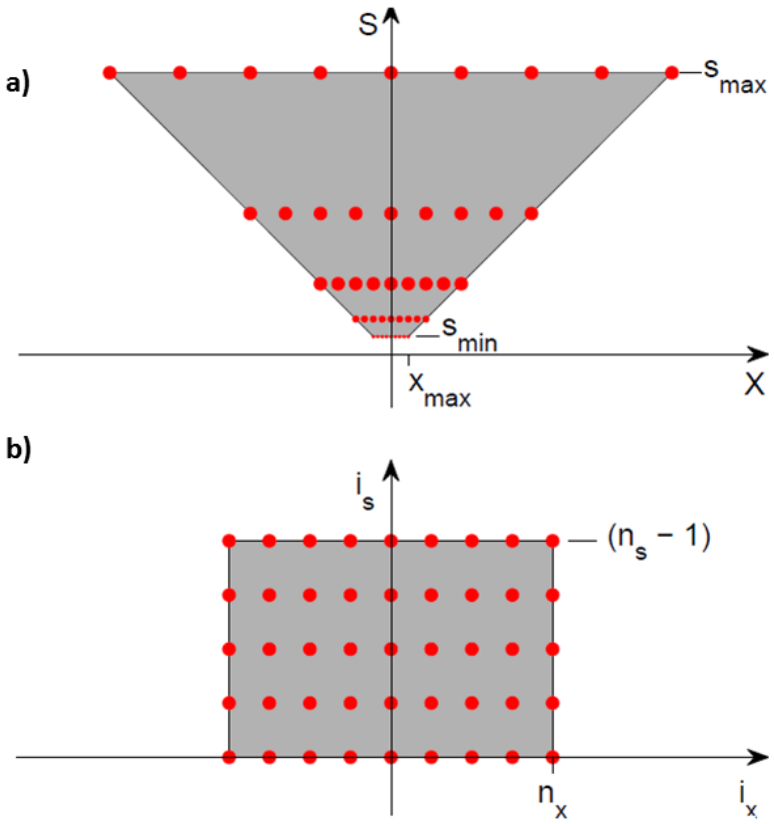

Figure 7: Under the assumption of Gabor filters and associated sampling for each scale at spatial intervals $\Delta x = s$, this graph depicts a subset of the resulting array of template units. Note that the sampling over x follows the sampling theorem. Note that there are no samples between the s axis and the line with slope 1 (when x and s are plotted in the same units). The center of the circles in the figure gives the s, x coordinates; the circles are icons symbolizing here the receptive fields. The figure shows a foveola of $\approx 26'$ and a total of ≈ 40 units (20 on each side of the center of the fovea). It also shows sampling intervals at the coarsest scale in the fovea (assumed to be around $2s = 21'$ (Marr et al.1980) which would span $\approx \pm 6$. Note that the size of a letter on the ophthalmologist's screen for $20/20$ vision is around $5'$. Since the inverted truncated pyramid a) has the same number of sample points at every scale, it maps perfectly onto a square array b) when x is replaced by $i_x = x/s$, i.e., the number of samples from the center. i_s is the scale band index. (s, x units are scaled as in Figure 6 for clarity). (From Poggio et al., 2014.) .

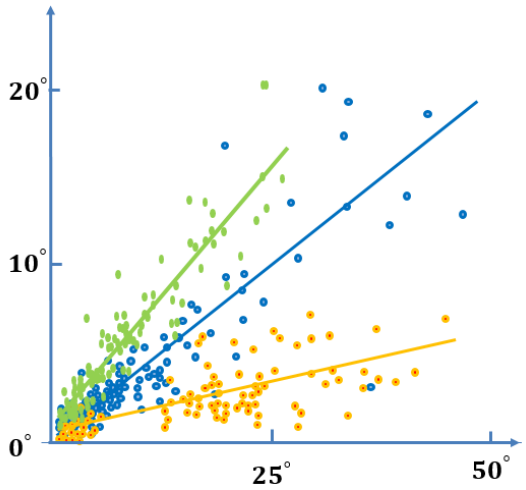

Figure 8: Data of Hubel and Wiesel for monkey V1 gives a slope for average RF diameter, relative to eccentricity, of $a = 0.05$ (from Hubel 1974). Data from other areas are similar but have higher slope (adapted from Freeman 2011 with original monkey data from Gattass 1981, Gattass 1988). (From Poggio et al., 2014.) .

1988) (shown in Figure 8) yield an estimate of the slope a for M in V1 (the slope of the line $s_{min}(x) = ax$). Hubel and Wiesel cite the slope of the average RF diameter in V1 as $a = 0.05$; Gattass quotes a slope of $a \approx 0.16$ (in both cases the combination of simple and complex cells may yield a biased estimate relative to the "true" slope of simple cells alone). Our model of an inverted truncated pyramid predicts (using an estimate of $a = 0.1$) from these estimates that the radius of the foveola (the bottom of the truncated pyramid) is $R = 1'20/0.1 \approx 13'$ with a full extent of $2R \approx 26'$ corresponding to about 40 cells separated by $40"$ each. The size of the fovea (the top of the truncated pyramid) would then have $2R \approx 6$ with 40 cells spaced $\approx 10''$ apart; see Figure 6. Each of the scale layers has the same number of units which is determined by the number of units in the fovea – that is, the number of units at the finest resolution. This remapping shows that S1 corresponds to a lattice of dimensions x, y, s, θ where the dimension sizes are different (but roughly the same for x, y); the topology is that of a cylinder with θ being periodic.

Notice that the lattice contains all the affine transformations - translations and scale – of the templates required for invariance within the inverted pyramid. The number is large but quite reasonable: $40 \times 40 = 1600$ per scale and per Gabor orientation for a total of around 50K transformed templates.
We have used data from the macaque together with data from human psychophysics. These estimates depend on the actual range of receptive field sizes and could easily be wrong by factors of 2. Our main goal is to provide a logical interpretation of future data and a ballpark estimate of relevant quantities.

4.3 Scale and position invariance in V1

Invariance is not provided directly by the array but by the pooling over it. Note that we are limiting ourselves in this section to the invariant recognition of isolated objects. The range of invariance in x is limited for each s by the slope of the lower bound of the inverted pyramid. The prediction is that the range of invariance Δx is proportional to the scale s,that is

$$\Delta x \approx n_x s \tag{11}$$

where n_x is the radius of the inverted pyramid, and is the same for all scales s. Thus small details (high frequencies) have a limited invariance range in x whereas larger details have larger invariance. n_x is obtained from the slope a of the cortical magnification factor as:

$$n_x = 1/a \tag{12}$$

The following rational for a so-called normative theory seems natural (and could easily be wrong):

- a) Evolution tries to optimize recognition from few labeled examples;
- b) as a consequence it has to optimize invariance;
- c) as a first step it has to optimize invariance to scale and translation (under constraint of non-infinite resources);
- d) therefore it develops in V1 multiple sizes receptive fields at each position with RF size increasing linearly with eccentricity, properties which are reflected in the architecture of the retina.

4.4 Tuning of cells in V1

Sections 3.5 and 3.4 describe the learning of templates through unsupervised experience of their transformations. The main example discussed consists of simple cells in V1. We can now add details to the example using the inverted pyramid architecture of this chapter. Our basic hypothesis is that the position and size (a Gaussian distribution) of each immature simple cells is set by development and corresponds to a node in the lattice of Figure 7 in x, y, s. Furthermore, Hebbian synapses on the simple cells will drive the tuning of the cell to be the eigenvector with the largest eigenvalue of the covariance of the input images. It can be shown (Poggio et al., 2013) that because of the Gaussian envelope at each site in the lattice and because of the statistics of natural images the tuning of each simple cell will converge to a Gabor functions with properties that match very closely the experimental data on the monkey, the cat and the mouse (see Figure 3). Since the arguments at the beginning of section 3.4 apply directly to this case, pooling over simple cells is equivalent to pooling over transformations of a template (a Gabor function with a specific orientation).

5 Stage 1: V2 and V4

5.1 Multistage pooling

As discussed in section 2.6, pooling in one-step over the whole s, x domain of Figure 6 suffers from the interference from clutter-induced fragments in any location in the inverted pyramid. A better strategy, as we will discuss in section 5.2, is to pool area by area (in the ventral stream), that is layer by layer (in the corresponding hierarchical architecture). The C cells activities in each layer of the hierarchy are sent to the memory or classification module at the top. After each pooling (C unit) stage (and possibly also after a dot product (S unit) stage), there is a downsampling of the array of units (in x, y and possibly in s) that follows from the low-pass-like effect of the operation. As a related remark, the templates in V2 and V4, according to the theory, should be "patches" of neural images at that level - possibly determined by the PCA-like learning described earlier.

Here we analyze the properties of a specific hypothetical strategy of downsampling in space by 2 at each stage. This choice is for simple and roughly consistent with biological data. It is easy to modify the results below by using different criteria for downsampling but the same logic. Pooling each unit over itself and its neighbors (thus a patch of radius 1) allows downsampling of the array in x by a factor of 2. We assume here that each combined S-C stage brings about a downsampling by 2 of each dimension of the x, y array. We call this process "decimation"; see Figure 9. Starting with V1, 4 stages of decimation (possibly at V1, V2, V4, TEO) reduce the number of units at each scale from ≈ 40 to ≈ 2 spanning $\approx 26'$ at the finest scale and ≈ 6 at the coarsest. Pooling over scale in a similar way may also decimate the array down to approximately one scale from V1 to IT. Neglecting orientations, in x, y, s the $30 \times 30 \times 6$ array of units may be reduced to just a few units in x, y and one in scale. This picture is consistent with the invariance found in IT cells (Serre et al. 2005). According to i-theory, different types of such units are needed, each for one of several templates at the top level. Other types of pooling and downsampling are conceivable such as pooling in space over each spatial frequency channel separately, possibly in different areas. Notice that the simple strategy of downsampling in space by 2 at each stage, together with our previous estimates of the dimensions of the inverted pyramid in V1, predicts that about 4-5 layers in the hierarchy are required to obtain almost full invariance to shift and scale before stage 2. If layers are identified with visual areas, there should be at least 4 areas from V1 to AIT: it is tempting to identify them with V1, V2, V4, PIT.

5.2 Predictions of crowding properties in the foveola and outside it (Bouma's law)

The pooling range is uniform across eccentricities in the plots of Figure 9. The spatial pooling range depends on the area: for V1 it is the sampling interval

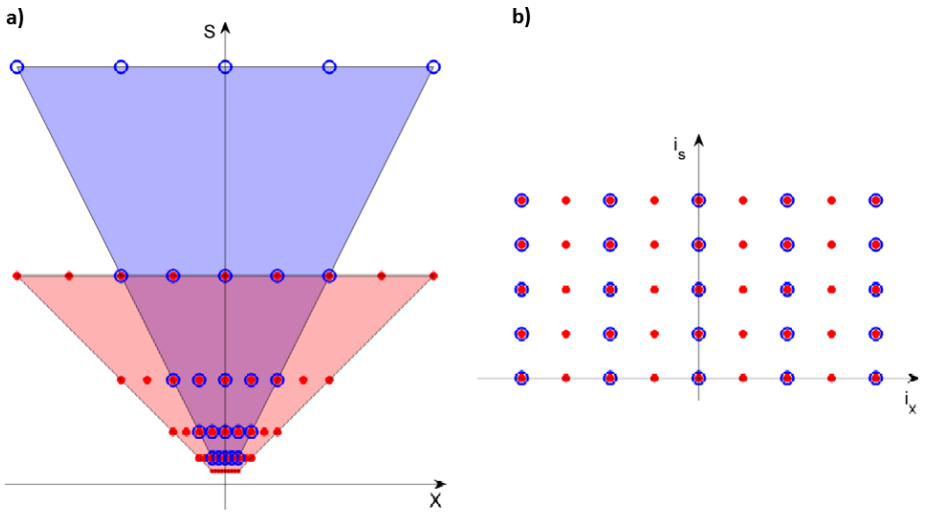

Figure 9: Pooling over 3×3 patches in x, y of the lattice of simple cells in V1 and subsampling decimates the lattice; if the lattice in x, y is 40 units, 4 steps (V1, V2, V4, PIT) of x pooling are sufficient to create cells that are 16 times larger than the largest in the fovea in V1 (probably around $21'$ at the coarsest scale), yielding cells with a RF diameter of up to ≈ 5. Each area in the fovea would see a doubling of size with corresponding doubling of the slopes at the border (before remapping to a cube lattice). The index of the units at position x and scale s is given by $i_x^s = 2^s i_x^1$. Simultaneous pooling over regions in S is possible. (From Poggio et al., 2014.)

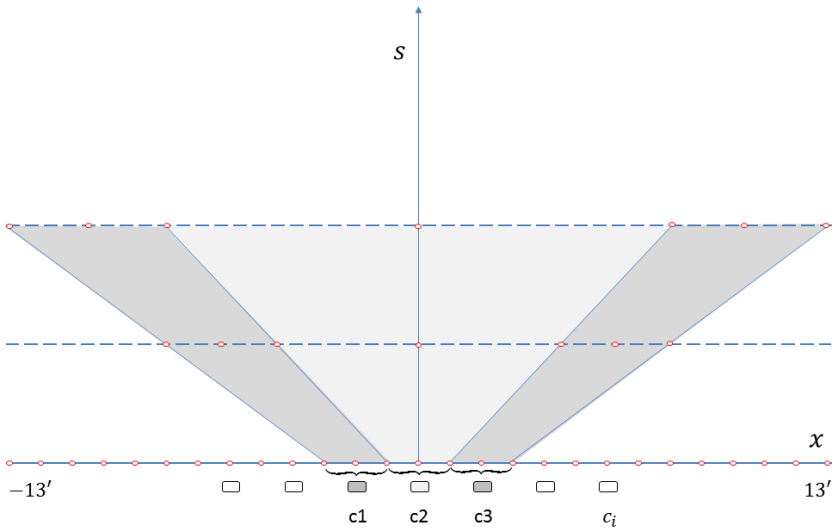

Figure 10: The diagram in x, s shows part of the inverted pyramid of simple units in V1. The shaded regions are pooling regions, each one corresponding to a complex cell (there are more such regions to cover the extent of the foveola (diameter of around $21'$): the ones at the left and right boundaries will be at the edges of the inverted pyramid.) Pooling shown here is spatial only (at each scale): it takes place around each simple units and its 2 neighbors in x, y (a 3×3 patch). Thus the critical spacing for crowding, at the highest resolution complex cells, is $\Delta x = 1'20"$ in V1, $\Delta x = 2'40"$ in V2, $\Delta x = 5'20"$ in V4 (assuming that the spacing among simple cells in V1 is $40"$).

between the red dots, for V2 it is the sampling interval between the blue dots: they should be roughly equivalent to the radius of the receptive field of the complex cells in V1 and V2 respectively. V4 is not shown but it is clear what is expected.

Figure 10 shows the regions of pooling corresponding to our assumptions. The inverted pyramid is split into sections, each one corresponding to pooling by a complex cell module. For now we consider only pooling in space and not scale. Figure 10 describes the situation for V1 but the same diagram also can be used for V2, V4 and PIT by taking into account the downsampling, the increase in size of the smallest cells and the doubling in slope of the lower boundaries of the inverted pyramid. It is graphically clear from the figure why the following criterion for pooling to remain interference-free from a flanking object, that is unaffected by clutter, seems reasonable: the target and the flanking distractor must be separated at least by the pooling range and thus by a complex cell

receptive field.
We consider two cases: a) the target is in the central section at some layer (say V2, for instance); b) the target is outside the foveola, that is at an eccentricity greater than 10′. The predictions are:

- a. Consider a small target, such as a $5'$ width letter, placed in the center of the fovea, activating the smallest simple cells at the bottom of the inverted pyramid. The smallest critical distance to avoid interference should be the size of a complex cell at the smallest scale, that is $\Delta x \approx 1'20"$ in V1 and $\Delta x \approx 2'40"$ in V2. If the letter is made larger, then the activation of the simple cells shifts to a larger scale (s in Figure 10) and since $\Delta x \approx s$, *the critical spacing is proportional to the size of the target*. It is remarkable that both these predictions match quite well Figure 10 in Levi and Carney, 2011.

- b. Usually the target is just large enough to be visible at that eccentricity (positive say). The target, such as a $8'$ width letter, placed at 1 degree eccentricity, is then in the section just above the lowest boundary of the inverted pyramid. The critical separation for avoiding crowding outside the foveola is

$$\Delta x \approx bx \tag{13}$$

  since the RF size of the complex cells increases linearly with eccentricity, with b depending on the cortical area responsible for the recognition signal (see Figure 9a). Thus the theory "predicts" Bouma's law , (Bouma, 1970) of crowding! (see also Peli and Tillman, 2008, Levi, 2008). The experimental value found by Bouma for crowding of $b \approx 0.4$ suggests that the area is V2 since this is the slope found by Gattass for the dependence of V2 RF size on eccentricity. Studies of "metameric" stimuli by Freeman and Simoncelli also implicated V2 in crowding and peripheral vision deficiencies, Freeman et al. 2011.

5.3 Scale and shift invariance dictates the architecture of the retina and of retinotopic cortex

It is interesting that, from the perspective of i-theory, the linear increase of RFs size with eccentricity, found in all primates, follows from the computational need of computing a scale and position invariant representation for novel images/objects. The theory also predicts the existence of a foveola and links its size to the slope of linear increase of receptive field size in V1. From this point of view, the eccentricity dependence of cone density in the retina as well as of the cortical magnification factor follow from the computational requirement of invariant recognition. The usual argument of limited resources (say number of fibers in the optical nerve) does not determine the shape of the inverted pyramid but only the size of the fovea at the bottom (and thereby of the total

number of cells). The inverted pyramid shape is independent of any bound on computational resources.

As we mentioned the estimate for the size of the foveola is quite small with a diameter of 26′ corresponding to about 40 simple cells (of each orientation) and about 50 cones in the retina. This is almost certainly a tradeoff between limited translation invariance - the inverted pyramid region - and the ability to correct for it by moving gaze, while scale invariance is fully available in the center of the fovea. Notice that a fovea with a 10 times larger diameter would lead to an optical nerve of the same size as the eye - making it impossible to move it.

This state of affairs means that there is a quite limited "field of vision" in a single glimpse. Most of an image of 5 by 5 degrees is seen at the coarsest resolution only, if fixation is in its center; at the highest resolution only a very small fraction of the image (up to $30'$) can be recognized, and an even smaller part of it can be recognized in a position invariant way (the number above are rough estimates). An "IP fragment" can be defined as the information captured from a single fixation of an image. Such a fragment is supported on a domain in the space x, s, contained in the inverted truncated pyramid of Figure 6. For normally-sized images and scenes with fixations well inside, the resulting IP-fragment will occupy most of the spatial and scale extent of the inverted pyramid. Consider now the fragment corresponding to an object to be stored in memory (during learning) and recognized (at run time). Consider the most favorable situation for the learning stage: the object is close to the observer so that both coarse and fine scales are present in its fragment. At run time then, the object can be recognized whenever it is closer or farther away (because of pooling). The important point is that a look at this and the other possible situations (at the learning stage) suggest that the matching should always weight more the finest available frequencies (bottom of the pyramid). This is the finding of Schynz (Schyns et al 2003). As implied by his work, top-down effects may modulate somewhat these weights (this could be done during pooling) depending on the task. Assume that such a fragment is stored in memory for each fixation of a novel image. Then, because of "large" and position-independent scale invariance, there is the following trade-off between learning and run-time recognition:

- if a new object is learned from a single fixation, recognition may require multiple fixations at run time to match the memory item (given its limited position invariance and unless fixation is set to be within the object).
- if a new object is learned from multiple fixations, with different fragments stored in memory each time, run time recognition will need a lower number of fixations (in expectation).

The fragments of an image stored in memory via multiple fixations could be organized into an egocentric map. Though the map may not be directly used for recognition, it is probably needed to plan fixations during the learning and especially during the recognition and verification stage (and thus indirectly

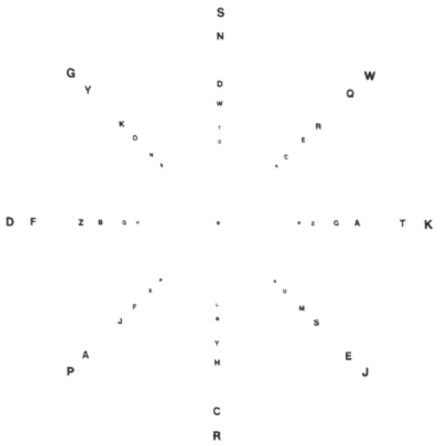

Figure 11: When fixating the central point, recognizability of a letter in the chart does not change under global scaling of the figure (from Anstis, 1974).

used for recognition in the spirit of minimal images and related recent work by Ullman and coworkers). Such a map may be related to Marr's $2 \cdot 1/2$ sketch, for which no neural correlate has been found as yet.

Thus, simultaneous invariance to translation and scale leads to an inverted, truncated pyramid as a model for the scale-space distribution of the receptive fields in V1; this is a new alternative, as far as we know, to the usual smooth empirical fit of the cortical magnification factor data. In particular, the linear dependency on eccentricity of receptive field sizes in V1 (and cones sampling in the retina) follows from the computational requirement of scale invariance. This picture contains other interesting properties: the range of shift invariance depends on scale; the size of the flat, high acuity foveal region, which we identify with the foveola can be inferred from the slope of the eccentricity dependent acuity. Existing neural data from macaque V1 suggest a foveola with a diameter of around $20'$ of arc. In a sense, scale invariance turns out to be more natural than shift invariance in this model (there is scale invariance over the full range at the fixation point). This is to be expected in organisms in which eye fixations can easily take care of shift transformations whereas more expensive motions of the whole body would be required to change the scale. I-theory predicts scale invariance at this level: this is exactly what Anstis (Anstis,1974) found (see Figure 11): a letter just recognizable at some eccentricity remains equally recognizable under scaling.

Physiology data suggest that AIT neurons are somewhat less tolerant to position changes of small stimuli (Op de Beeck et al. 2001), also described in (Di Carlo et al. 2003). A comparison across studies suggests that position tolerance is roughly proportional to stimulus size (Di Carlo et al. 2003). If we assume that some IT neurons effectively pool over "all" positions and scales i-theory in fact

expects that their invariant receptive field (over which consistent ranking of stimuli is maintained) should be smaller for higher spatial frequency patterns than with low frequency ones. There is evidence of attentional suppression of non-attended visual areas in V4. From the perspective of i-theory, it seems natural that top-down signals may be able to control the extent of pooling, or the pooling stage, used in computing a signature from a region of the visual field, in order to minimize clutter interference.

In summary, I-theory provides explanations and computational justifications for several known properties of retinotopic cortex. It also makes a few predictions that are still waiting for an experimental test:

- There is an inverted pyramid of simple cells size and positions with parameters specified in the text, including linear slope of the lower boundaries. The predicted pyramid is consistent with available data. More precise measurements in the region of the foveola could decide between the usual empirical fits and our predictions.

- Anstis did not take measurements close to the minimum letter size – which is around $5'$ for $20/20$ vision. I-theory predicts that if there is a range of receptive fields in V1 between s_{min} and s_{max} in the fovea then there is a finite range of scaling between s_{min} and s_{max} under which recognition is maintained (see Poggio et al. 2014). It is obvious that looking at the image from an increasing distance will at some point make it unrecognizable; it is somewhat less obvious that getting too close will also make it unrecognizable (this phenomenon was found in Ullman's minimal images; Ullman, personal comm.)

- Consider the experimental use of images such as novel letters (never seen before) of appropriate sizes that are bandpass filtered (with the Gabor-like filters assumed for V1). The predictions - because of the pooling over the whole inverted pyramid (done between V1, V2 and V4) – is that for a new presentation there will be psychophysically scale invariance for all frequencies between s_{min} and s_{max}; there is shift invariance that increases linearly with spatial wavelength and is at any spatial frequency at least between x_{min} and x_{max} (the bottom edge of the truncated pyramid).

- I-theory predicts a flat region of constant maximum resolution - that we called foveola. Its size determines the slope of the lower border of the pyramid. Since the slope can be estimated relatively easily from existing data, our prediction for the linear size of the foveola is around 40 minutes of arc, corresponding to about 30 simple cells of the smallest size (assumed to be $\approx 1'20''$ of arc). Notice our definition of the fovea is in terms of the set of all scaled versions of the foveola between s_{min} and s_{max} spanning about 6 degrees of visual angle.

- I-theory explains crowding effects in terms of clutter interference in the pooling stage (see also Balas et al., 2009). It predicts Bouma's law and its linear dependency on eccentricity (Bouma,1970).

- Since Bouma's constant has a value of about 0.4 (see also Freeman et al. 2011), our theory requires that a signature that is interference-free from clutter at the level of V2 is critical for recognition. This is consistent with i-theory independent requirement (Anselmi et al. 2014, Anselmi et al. 2013) that signals associated with image patches of increasing size from different visual areas must be able to access memory and classification stages separately. The requirement follows from the need of recognizing "parts and wholes" in an image and avoid clutter for small objects. The V2 signal could directly or indirectly (via IT or V4 and IT) reach memory and classification.

- The angular size of the fovea remains the same at all stages of a hierarchical architecture (V1, V2, V4...), but the number of units per unit of visual angle decreases and the slope increases because the associated x increases (see Figure 9).

- The theory predicts crowding in the foveola (very close to fixation) but with very small Δx (see equation 10) that depends on the size of the (small) objects rather than eccentricity. For objects smaller than $20'$ in diameter the prediction (to be tested) is $\Delta x \approx 3' - 4'$, assuming that the main effect of clutter is at the smallest of the five channels of simple cells and in V2.

5.4 Tuning of "simple" cells in V2 and V4

It is difficult to make clear predictions about V2 and V4, because several options are theoretically possible. A simple scenario is as follows. There is a x, y, s lattice for V2 (and another for V3) simple cells as shown in Figure 6. The tuning of each simple cell - a point in the x, y, s lattice – is determined by the top PCAs computed on the neural activity of the complex cells in V1, seen through a Gaussian window that includes $\approx 3 \times 3$ complex cells in x, y and a set of orientations for each position and scale. It is likely - and supported by preliminary experiments (Poggio et al., 2013) - that some of the PCA computed in this way over a large number of natural images can lead to cell tunings similar to measurements in V4.

6 Stage 2 in IT: class-specific approximate invariance

6.1 From generic templates to class-specific tuning

As discussed in section 2.5, approximate invariance for transformations beyond the affine group requires highly tuned templates, therefore highly tuned simple cells, probably at a level in the hierarchy corresponding to AIT. According to the considerations of section 2.6 this is expected to take place in higher visual areas of the hierarchy. In fact, the same localization condition Equation 4 suggests Gabor-like templates for generic images in the first layers of a hierarchical architecture and specific tuned templates for the last stages of the hierarchy, since class specific modules are needed, one for each class and each containing highly specific templates,that is highly tuned cells. This is consistent with the architecture of the ventral stream and the the existence of class-specific modules in primate cortex such as a face module and a body module (Tsao, 2003, Leibo et al. 2011a, Kanwisher, 2010, Downing and Jiang 2001). We saw in section 2.6 that areas in the hierarchy up to V4 and/or PIT provide signatures for larger parts or full objects. Thus we expect

- that the inputs to the class-specific modules are scale and shift invariant
- that the class-specific templates are "large". For instance in the case of faces, templates should cover significant regions of the face. Notice that only large templates support pose invariance: the image of an isolated eye does not change much under rotations in depth of the face!

6.2 Development of class-specific and object-specific modules

A conjecture emerging from i-theory offers an interesting perspective (Leibo et al. 2014) on AIT. For transformations that are not affine transformations in 2D (we assume that 3D information is not available to the visual system or used by it, which may not always be true), an invariant representation cannot be computed from a single view of a novel object because the information available is not sufficient. What is lacking is the 3D structure and material properties of the object: thus exact invariance to rotations in depth or to changes in the direction or spectrum of the illuminant cannot be obtained. However, as i-theory shows, approximate invariance to smooth non-group transformations can still be achieved in several cases (but not always) using the same HW module. The reason this will often approximately work is because it effectively exploits prior knowledge of how similar objects transform. The image-to-image transformations caused by a rotation in depth are not the same for two objects with different 3D structures. However, objects that belong to an object class where all the objects have similar 3D structure transform their 2D appearance in (approximately) the same way. This commonality is exploited by a HW module to transfer the invariance learned from (unsupervised) experience with template

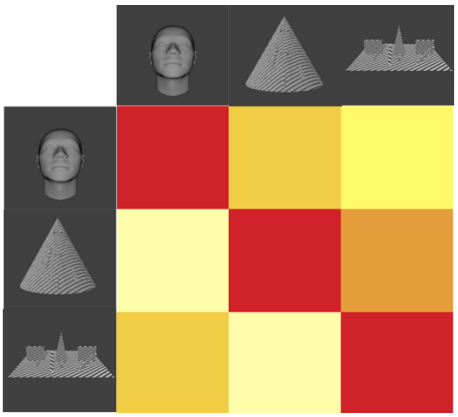

Figure 12: Class-specific transfer of depth-rotation invariance for images from three classes (faces, A, cylinders, B and composed, C). The left column of the matrix shows the results of the test for invariance for a random image of a face (A) in different poses w.r.t. 3D rotation using 3D rotated templates from A,B,C; similarly the middle and the right column shows the invariance results for class B and C tested on rotated templates of A,B,C respectively. The colors in the matrix show the maximum invariance range (degrees of rotation away from the frontal view). Only the diagonal values of the matrix (train A - test A, train B - test B, train C- test C) show an improvement of the view-based model over the pixel representation. That is, only when the test images transform similarly to the templates is there any benefit from pooling (Leibo et al. 2014).

objects to novel objects seen only from a single example view. This is effectively our definition of an object class: a class of objects such that the transformation for a specific object can be approximately inferred from how other objects in the class transform. The necessary condition for this to hold is that the 3D shape is similar between any two objects in the class. The simulation in Figure 12 shows that HW-modules tuned to templates from the same class of the (always novel) test objects provide a signature that tolerates substantial viewpoint changes (plots on the diagonal); it also shows the deleterious effect of using templates from the wrong class (plots off the diagonal). There are of course several other class-specific transformations besides depth-rotation, such as face expression and body pose transformations.

In an interesting conjecture, (Leibo et al. 2014) argue that the visual system is continuously and automatically clustering objects and their transformations - observed in an unsupervised way - in class-specific modules. Images of an object and of its transformations correspond to a orbit Γ_k. New images are added to an existing module only if their transformation are well predicted by it. If no module can be found with this property the new orbit will be the seed of a new object cluster/module.

For the special case of rotation in depth, (Leibo et al. 2014), ran a simula-

tion using 3D modelling / rendering software to obtain the orbits of objects for which there exist 3D models. Faces had the highest degree of clustering of any naturalistic category - unsurprising since recognizability likely influenced face evolution. A set of chair objects had broad clustering, implying that little invariance would be obtained from a chair-specific region. A set of synthetic "wire" objects, very similar to the "paperclip" objects used in several classic experiments on view-based recognition e.g. (Bar et al. 2008, Logothetis et al. 1994, Logothetis et al. 1995) were found to have the smallest index of clusterability: experience with familiar wire objects does not transfer to new wire objects (because the 3D structure is different for each individual paperclip object).

It is instructive to consider the limit case of object classes that consist of single objects - such as individual paperclips. If the object is observed under rotation several frames are memorized as transformations of a single template (identity is implicitly assumed to be conserved by a Foldiak-like rule, as long as there is continuity in time of the transformation). The usual HW module pooling over them will allow view-independent recognition of the specific object. A few comments:

1. remarkably, the HW module described above for class-specific transformations – when restricted to multiple-views, single-object – is equivalent [6] to the Edelman-Poggio model for view invariance (Edelman and Poggio 1990);

2. the class-specific module is also effectively a "gate": in addition to providing a degree of invariance it also performs a template matching operation with templates that can effectively "block" images of other object classes. This gating effect may be important for the system of face patches discovered by Tsao and Freiwald and it is especially obvious in the case of a single object module;

3. from the point of view of evolution, the use of the HW module for class-specific invariances can be seen as a natural extension from its role in single-objects view invariance. The latter case is computationally less interesting, since it implements effectively a look-up table, albeit with interpolation power. The earlier case is more interesting since it allows generalization from a single view of a novel object. It also represent a clear case of transfer of learning.

6.3 Domain-specific regions in the ventral stream

As discussed by (Leibo et al. 2014), there are other domain-specific regions in the ventral stream besides faces and bodies. It is possible that additional

[6] In Edelman and Poggio 1990 the similarity operation was the Gaussian of a distance - instead of the dot product required by i-theory. Notice that for normalized vectors, l_2 norms and dot products are equivalent.

regions for less-common or less transformation-compatible object classes will appear with higher resolution imaging techniques. One example may be the fruit area, discovered in macaques with high-field fMRI (Ku et al. 2011). Others include the body area and the Lateral Occipital Complex (LOC) which according to recent data (Malach et al. 1995) is not really a dedicated region for general object processing but a heterogeneous area of cortex containing many domain-specific regions too small to be detected with the resolution of fMRI. The Visual Word Form Area (VWFA) (Cohen et al. 2000) seems to represent printed words. In addition to the generic transformations that apply to all objects, printed words undergo several nongeneric transformations that never occur with other objects. For instance, our reading is rather invariant to font transformations and can deal with hand-written text. Thus, VWFA is well-accounted for by the invariance hypothesis. Words are frequently-viewed stimuli which undergo class-specific transformations.

The justification - really a prediction! - by i-theory for domain-specific regions in cortex is different from other proposals. However, it is in complementary w.r.t. some of them, rather than exclusive. For instance, it would make sense that the clustering depends not only on the index of compatibility but also on the relative frequency of each object class. The conjecture claims a) that transformation compatibility is the critical factor driving the development of domain-specific regions, and b) that there are separate modules for object classes that transform differently from one another.

6.4 Tuning of "simple" cells in IT

In the case of "simple" neurons in the AL face patch (Freiwald et al 2010 and Leibo et al, in preparation), exposure to several different faces – each one generating several images corresponding to different rotations in depth – yields a set of views with a covariance function which has eigenvectors (PCs) that are either even or odd functions (because faces are bilaterally symmetric; par 5.4.1, pg. 110 Magic Material 2013) .

The Class-specific theorem together with the Spectral pooling proposition suggests that square pooling (over these face PCs provides approximate invariance to rotations in depth. The full argument goes as follows. Rotations in depth of a face around a certain viewpoint - say $\theta = \theta^0$ – can be approximated well by linear transformations (by transformations $g \in GL(2)$). The HW algorithm can then provide invariance around $\theta = \theta^0$. Finally, if different sets of "simple" cells are plastic at somewhat different times, exposure to a partly different set of faces yields different eigenvectors summarizing different sets of faces. The different sets of faces play the role of different object templates in the standard theory.

The limit case of object classes that consist of single objects is important to understand the functional architecture of most of IT. If an object is observed under transformations, several images of it can be memorized and linked together by continuity in time of the transformation. As we mentioned, the usual HW module pooling over them will allow view-independent recognition of the

specific object. Since this is equivalent to the Edelman-Poggio model for view invariance (Edelman and Poggio 1990) there is physiological support for this proposal (see Logothetis, Pauls and Poggio, 1995; Logohetis and Sheinberg, 1996; Stryker, 1991).

6.5 Mirror symmetric tuning in the face patches and pooling over PCs

The theory then offers a direct interpretation of the Tsao-Freiwald data (see Freiwald et and Tsao 2010, Freiwald et al. 2009) on the face patch system. The most posterior patches (ML/MF) provide a view and identity specific input to the anterior patch AL where most neurons show tuning which is an even function of the rotation angle around the vertical axis. AM, which receives inputs from AL, is identity-specific and view-invariant. The puzzling aspect of this data is the mirror symmetric tuning in AL: why does this appear in the course of a computation that leads to view-invariance? According to the theory the result should be expected if AL contains "simple" cells that are tuned by a synaptic Hebb-like Oja rule and the output of the cells is roughly a squaring nonlinearity as required by the Spectral pooling proposition. In this interpretation, cells in AM pool over several of the squared eigenvector filters to obtain invariant second moments (see Figure 13). Detailed models from V1 to AM show properties that are consistent with the data and also perform well in invariant face recognition (Liao et al. 2013, Leibo et al. 2011b, Leibo2013, Mutch et al. 2010).

7 Discussion

Several different levels of understanding. I-theory is at several different levels addressing the computational goal of the ventral stream, the algorithms used, down to the architecture of visual cortex, its hierarchical architecture and the neural circuits underlying tuning of cells. This is unlike most other models or theories.

Predictions. From the point of view of neuroscience, the theory makes a number of predictions, some obvious, some less so. One of the main predictions is that simple and complex cells should be found in all visual and auditory areas, not only in V1. Our definition of simple cells and complex cells is different from the traditional ones used by physiologists; for example, we propose a broader interpretation of complex cells, which in the theory represent invariant measurements associated with histograms of the outputs of simple cells or of moments of it. The theory implies that, under some conditions, exact or approximate invariance to all geometric image transformations can be learned, either during development or in adult life. It is, however, also consistent with the possibility that basic invariances may be genetically encoded by evolution and possibly refined and maintained by unsupervised visual experience. A

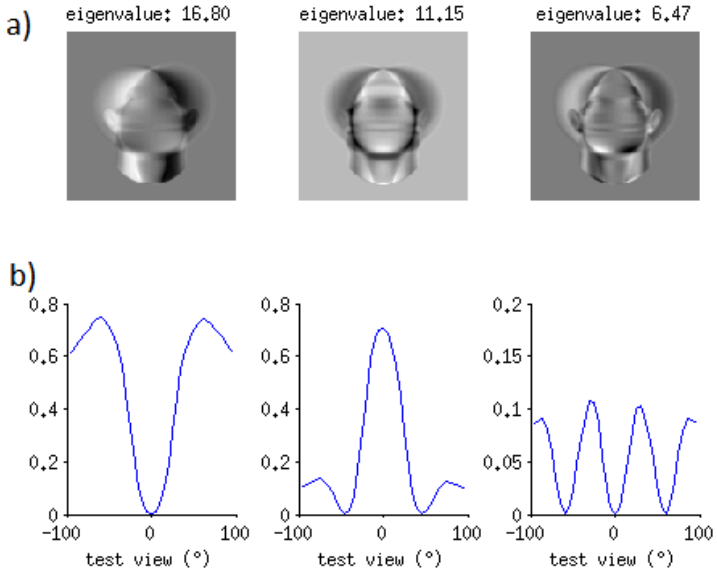

Figure 13: Face identity is represented in the macaque face patches (Freiwald and Tsao, 2010). Neurons in the middle areas of the ventral stream face patch (middle lateral and fundus (ML, MF)) are view specific, while those in the most anterior (anterior medial patch (AM)) are view invariant. Neurons in an intermediate area (anterior lateral patch (AL)) respond similarly to mirror-symmetric views. In i-theory view invariance is obtained by pooling over "simple" neurons whose tuning corresponds to the PCAs of a set of faces previously experienced each under a range of poses. Due to the bilateral symmetry of faces, the eigenvectors of the associated covariance matrix are even or odd. This is shown in a) where the first 3 PCAs of set of grey-level faces under different poses are plotted: the same symmetry arguments apply to "neural" images of faces. b) shows the response of 3 model AL units to a face stimulus as a function of pose under different poses (Leibo et al. 2014).

single cell model for simple complex cells follows from the theory as an interesting possibility. I-theory also makes predictions about the architecture of the ventral stream:

- the output of V2, V4, PIT should access memory either via connections that bypass higher areas or indirectly via equivalent neurons in higher areas (because of the argument in a previous section about clutter).
- areas V1, V2, V4 and possibly PIT are mainly dedicated to compute signatures that are invariant to translation, scale and their combinations - as experienced in past visual experience.
- IT is a complex of parallel class-specifc modules for a large number of large and small object classes. These modules receive position and scale invariant inputs (invariance in the inputs greatly facilitates unsupervised learning of class specific transformations). We recall that, from the perspective of the theory, the data of Logothetis et al. 1995 concern single object modules and strongly support the prediction that exposure to a transformation lead to neuronal tuning to several "frames" of it.

Object-based vs 3D vs view-based recognition. We should mention here an old controversy about whether visual recognition is based on views or on 3D primitive shapes called geons. In the light of i-theory image views retain the main role but ideas related to 3D shape may also be valid. The psychophysical experiments of Edelman and Buelthoff concluded that generalization for rotations in depth was limited to a few degrees ($\approx \pm$ 30 degrees) around a view (independently of whether 2D or 3D information was provided to the human observer (psychophysics in monkey (Logothetis et al 1994, 1995) yielded similar results). The experiments were carried out using "paperclip" objects with random 3D structure (or similar but smoother objects). For this type of objects, class-specific learning is impossible (they do not satisfy the second condition in the class-specific theorem) and thus i-theory predicts the result obtained by Edelman and Buelthoff. For other objects, however, such as faces, the generalization that can be achieved from a single view by i-theory can span a much larger range than ± 30 degrees, effectively exploiting 3D-like information from templates of the same class.

Genes or learning. I-theory shows how the tuning of the "simple" cells in V1 and other areas could be learned in an unsupervised way. It is possible however that the tuning - or better the ability to quickly develop it in interaction with the environment - may have been partly compiled during evolution into the genes [7]. Notice that this hypothesis implies that most of the times the specific function is not be fully encoded in the genes: genes facilitate learning but

[7] If a function learned by an individual represents a significant evolutionary advantage we could expect that aspects of the learning the specific function may be encoded in the genes, since an individual who learns more quickly has a significant advantage. In other words, the hypothesis implies a mix of nature and nurture in most competencies that depend on learning from the environment (like perception). This is an interest-

do not replace it completely. It has to be expected then in the "nature vs nurture debate" that usually nature needs nurture and nurture is made easier by nature.

Computational structure of the HW module. The HW module computes the CDF of $\langle I, g_i t^k \rangle$ over all $g_i \in G$. The computation consists of

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + h\Delta) \tag{14}$$

with $h = 0, ..., H$ and $k = 1, ..., K$; the main forms of the nonlinearity σ are either a threshold function or a power $n = 1, ..., \infty$ of its argument. Several known networks are special cases of this module. One interesting case is when G is the translation group and $\sigma(\cdot) = || \cdot ||^2$: then the equation is equivalent (for $H = 0$) to a unit in a convolutional network with max pooling. In another noteworthy case (we always assume that I and t^k are normalized) the equation is very similar to the RBF network proposed by Edelman and Poggio (Poggio and Edelman1990) for view classification. In this spirit, note that the equation for a unit in a convolutional network is

$$\frac{1}{|G|} \sum_{i=1}^{|G|} c_i \sigma(\langle I, g_i t \rangle + h\Delta) \tag{15}$$

where I is the input vector, c_i, t, Δ are parameters to be learned, in supervised mode, from labeled data and $g_i t(x) = t(x - i\delta_x)$. Thus units in convolutional network could learn to become units of the i-theory (by learning $c_i = 1$) but only when G is the translation group (in the i-theory G is the full affine group for the first layers and can be a non group such as the transformation induced by rotations in depth).

Relations to Deep Learning networks. The best performing deep learning networks - a new name for multilayer perceptrons (MLPs) - have convolutional layers as well as densely connected layers. I-theory applies to the convolutional but not the densely connected, classification stage. Historically, hardwired invariance to translation was first introduced in the Neocognitron by Fukushima and later in LeNet (LeCun et al 1995) and in HMAX (Riesenhuber

ing implication of the "Baldwin effect" - a scenario in which a character or trait change occurring in an organism as a result of its interaction with its environment becomes gradually assimilated into its developmental genetic or epigenetic repertoire (Simpson, 1953; Newman, 2002). In the words of Daniel Dennett, "Thanks to the Baldwin effect, species can be said to pretest the efficacy of particular different designs by phenotypic (individual) exploration of the space of nearby possibilities. If a particularly winning setting is thereby discovered, this discovery will create a new selection pressure: organisms that are closer in the adaptive landscape to that discovery will have a clear advantage over those more distant." (p. 69, quoting Dennett 1991).

2000; HMAX had also invariance to scale). These architectures are early examples of convolutional networks. I-theory provides a general theory for them [8] that also offers two significant algorithmic and architectural extensions: a) it ensures, within the same algorithm, invariances to other groups beyond translation and, in an approximate way, to certain non-group transformations; b) it provides a way to learn arbitrary invariances from unsupervised learning.

Invariance in 2D and 3D vision. We have assumed here that "images" as well as templates are in 2D. This is the case if possible sources of 3D information such as stereopsis and or motion are eliminated. Interestingly, it seems that stereopsis does not facilitate recognition, suggesting that 3D information, even when available, is not used by the human visual system (see Bricolo 1996) [9].

Relations to the scattering transform. There are connections between the scattering transform and i-theory but also several differences. There is no obvious correspondence between operations in the scattering transform and simple+complex cells in the ventral stream unlike convolutional networks and i-theory networks. I-theory provides an algorithm in which invariances are learned from unsupervised experience of transformations of a random set of objects/images; in the scattering transform invariances are hardwired. I-theory proves that Gabor-like templates are optimal for simultaneous invariance to scale and shift and that such invariance requires a multi-resolution inverted and truncated pyramid which turns out to be reflected in the architecture of the visual cortex starting with eccentricity dependent organization of the retina; in the scattering transform, Gabor wavelets are assumed at the start.

Explicit or implicit gating of object classes. The second stage of the recognition architecture consists of a large set of object-class specific modules of which probably the most important is the face system. It is natural to think that signals from lower areas should be gated, in order to route access only to the appropriate module. In fact, Tsao (Tsao and Livingstone 2008) postulated a gate mechanism for the network of face patches. The structure of the modules however suggests that the module themselves provides automatically a gating function even if their primary computational function is invariance. This is especially clear in the case of the module associated with a single object (the object class consists of a single object as in the case of a paperclip). The input to the module is subject to dot products with each of the stored views of the object: if none matches well enough the output of the module will be close to

[8] In the case of the translation group the HW module (see Equation 1) consists of (non-linear) pooling of the convolution of the image with a template.

[9] This hypothesis should however be checked further since i-theory implies that if 3D information is available, rotation in depth is a group and therefore generalization from a single view could be available simply by having stored 3D templates of a few arbitrary objects and their 3D transformations. This is not what psychophysics (for instance on the paperclips) shows; however, the mathematical claim of perfect invariance is only true in the absence of self-occlusions, a clearly unrealistic assumption for most objects.

zero, effectively gating off the signal and switching off subsequent stages of processing.

Invariance to X and estimation of X. Our description of i-theory focuses on the problem of recognition as estimating identity or category invariantly to a transformation X - such as translation or scale or pose. Often however, the complementary problem, of estimating X, for instance pose, is also important. The same neural population may be able to support both computations and multiplex the representations of their outcome as shown in IT recordings (Hung et al 2005) and model simulations (Serre at al. 2005). As human observers, we are certainly able to estimate position, rotation, illumination of an object without eye movements. HW modules pooling over the same units in different way - pooling over identities for each pose or pooling over pose for each identity - can provide the different types of information using the same "simple" cells and different "complex" cells. Anselmi et al. (2013, fig 45) show simulations of recognizing a specific body invariantly to pose and estimating pose-out of a set of 32 possibilities-of a body invariantly to the identity.

PCAs vs ICAs. Independent Component Analysis (ICA) (Hyvrinen and Oja 2000) and similar unsupervised mechanisms describe plasticity rules similar to the basic Oja flow analyzed in this paper. They can generate Gabor-like receptive fields and they may not need the assumption of different sizes of Gaussian distributions of LGN synapses. We used PCA simply because its properties are easier to analyze and should be indicative of the properties of similar Hebbian-like mechanisms. Parsing a scene. Full parsing of a scene cannot be done in a single feedforward processing step in the ventral stream. It requires task-dependent top-down control, in general multiple fixations and therefore observation times longer than ≈ 100 msec. This follows also from the limited high resolution region of the inverted pyramid model of the visual system, that the theory predicts as a consequence of simultaneous invariance to shift and scale. In any case, full parsing of a scene is beyond what a purely feedforward model can provide.

Feedforward and feedback. We have reviewed a forward theory of recognition and some of the related evidence. I-theory does not address top-down or recurrent or horizontal connectivity and their computational role. It makes however easier to consider plausible hypothesis. The inverted pyramid architecture that follows from scale and position invariance requires for everyday vision a tight loop between different fixations in which an efficient control module drives eye movements by combining task requirements with memory access. Within a single fixation, however, the space-scale inverted pyramid cannot be shifted in space. What could be controlled in a feedback mode are parameters of pooling, including the choice of which scales to use depending on the results of classification or memory access. The most obvious limitation of feedforward architectures is recognition in clutter and the most obvious way around the problem is the attentional masking of large parts of the image under top-down control.

More in general, a realistic implementation of the present theory requires top-down control signals and circuits, supervising learning and possibly fetching signatures from different areas and at different locations in a task-dependent way. An even more interesting hypothesis is that backprojections update local signatures at lower levels depending on the scene class currently detected at the top (an operation similar to the top-down pass of Ullman (Borestein and Ullman 2008). In summary, the output of the feedforward pass is used to retrieve labels and routines associated with the image; backprojections may implement an attentional focus of processing to reduce clutter effects and also to run visual routines (Serre et al. 2005) at various levels of the hierarchy.

Motion helps learning isolated templates. Ideally templates and their transformations should be learned without clutter. It can be argued that if the background changes between transformed images of the same template the averaging effect intrinsic to pooling will mostly "average out" the effect of clutter during the unsupervised learning stage. Though this is correct and we have computer simulations that provide empirical support to the argument, it is interesting to speculate that motion could provide a simple way to eliminate most of the background. Sensitivity to motion is one of the earliest visual computations to appear in the course of evolution and one of the most primitive. Stationary images on the retina tend to fade away. Detection of relative movement is a strong perceptual cue in primate vision as well as in insects vision, probably with similar normalization-like mechanisms (Heeger 1992, Poggio and Reichardt 1973). Motion induced by the transformation of a template may then serve two important roles:

- to bind together images of the same template while transforming: continuity of motion is implicitly used to ensure that identity is preserved;
- to eliminate background and clutter by effectively using relative motion.

The required mechanisms are probably available in the retina and early visual cortex.

I-theory. The theory that guided our review of computational aspects of the ventral stream cuts across levels of analysis (Marr 1976). Some of the existing models between neuroscience and machine learning, such as HMAX (Riesenhuber and Poggio 2000, Mutch and Lowe 2006, Serre et al 2007) and other convolutional neural networks (Fukushima 1980, LeCun 1989, LeCun 1995, LeCun 2004), are special cases of the theory. Despite significant advances in sensory neuroscience over the last five decades, a true understanding of the basic functions of the ventral stream in visual cortex has proven to be elusive. Thus it is interesting that the theory used in this paper follows from a novel hypothesis about the main computational function of the ventral stream: the representation of new objects/images in terms of a signature which is invariant to transformations learned during visual experience, thereby allowing recognition from very few labeled examples—in the limit, just one. This

view of the cortex may also represent a novel theoretical framework for the next major challenge in learning theory beyond the supervised learning setting which is now relatively mature: the problem of representation learning, formulated here as the unsupervised learning of invariant representations that significantly reduce the sample complexity of the supervised learning stage.

Acknowledgments Thanks to Danny Harari, Leyla Isik, Lorenzo Rosasco and especially to Gabriel Kreiman.

References

Abdel-Hamid, O. and Mohamed, A. and Jiang, H. and Penn, G., (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, ICASSP 4277-4280.

Y. S. Abu-Mostafa (1993). Hints and the VC dimension. Neural Computation, 5(2):278-288.

Adelson E. and Bergen J. (1985). Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A, 2(2):284-299.

Anselmi, F. Leibo, J.Z., Rosasco, L., Mutch, J., Tacchetti, A., and Poggio, T.(2014). Unsupervised learning of invariant representations with low sample complexity: the magic of sensory cortex or a new framework for machine learning?. CBMM Memo No. 001. arXiv:1311.4158v5.

Anselmi F., J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio (2013) "Magic Materials: a theory of deep hierarchical architectures for learning sensory representations", CBCL paper, Massachusetts Institute of Technology, Cambridge, MA.

Anstis, S.(1974). A chart demonstrating variations in acuity with retinal position. Vision Research, (14):589-592.

Balas B., L. Nakano, and R. Rosenholtz (2009) A summary-statistic representation in peripheral vision explains visual crowding, Journal of Vision.

Bar, M., Aminoff, E., and Schacter, D. L. (2008). Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. The Journal of Neuroscience, 28(34):8539-8544.

Blender.org, "Blender 2.6," 2013. 78, 108.

Bouma, H., (1970). Interaction effects in parafoveal letter recognition. Nature, (226):177-178.

Borenstein E., Ullman S., "Combined Top-Down/Bottom-Up Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30, no. 12, pp. 2109-2125.

Bricolo, E. (1996). Ph.D. Thesis, BCS, MIT: On the Representation of Novel Objects: Human Psychophysics,Monkey Physiology and Computational Models.

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

Cowey, A. and Rolls, E. T.(1974). Human cortical magnification factor and its relation to visual acuity. Experimental Brain Research, (21):447-454.

De Beek, H. O. and Vogels, R. (2001). Spatial sensitivity of macaque inferior temporal neurons. Journal of Comparative Neurology, (426):500-518.

Dennett, D. (2003), The Baldwin Effect, a Crane, not a Skyhook in : Weber, Bruce H.; Depew, David J. (2003). Evolution and learning: The Baldwin effect reconsidered. Cambridge, MA: MIT Press. pp. 69-106.

DiCarlo, J. J. and Maunsell, J.H. R. (2003), Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position, J Neurophysiol 89, 3264-3278.

Digimation.com. Digimation archive.

Donoho, S. P. (1989) Uncertainty principles and signal recovery. SIAM J. Appl. Math, 49(3):906-931.

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

Blthoff H.H., Edelman S.(1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition PNAS vol. 89 no. 1.

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

Foldiak, P. (1991). Learning invariance from transformation sequences. Neural Computation, 3(2):194-200.

Freeman, J. and Simoncelli, E. (2011). Metamers of the ventral stream. Nature Neuroscience, 14:11951201.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model

for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193 202.

Freiwald W.A. and Tsao D.Y. (2010) Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. Science 330(6005): 845-851.

Freiwald WA, Tsao DY, Livingstone M.S. (2009) A face feature space in the macaque temporal lobe. Nat Neurosci. 2009.

Gabor D. Theory of communication (1946). Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering, 93(26):429-457.

Gallant J., Braun J., and Essen D. V.(1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science.

Gallant J., Connor C., Rakshit S., Lewis J., and Van Essen D (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. Journal of Neurophysiology, 76:2718-2739.

Gattass, R., Gross, C. G., and Sandell, J. H. (1981). Visual topography of v2 in the macaque. The Journal of comparative neurology, (201):519-39.

Gattass, R., Sousa, A., and Gross, C. G. (1988). Visuotopic organization and extent of v3 and v4 of the macaque. J. Neurosci., (6):1831-1845.

Gauthier, I. and Jarr, M. (1997). Becoming a" greeble" expert: Exploring mechanisms for face recognition. Vision Research, 37(12):1673-1682.

Google (2014) Going deeper with convolutions http://arxiv.org/abs/1409.4842.

Heeger D.J. 1992. Normalization of cell responses in cat striate cortex. Vis. Neurosci.1992;9:181-197.

Hebb, D. O. (1949). The Organization of Behaviour: A Neuropsychological Theory, Wiley.

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

Heppes A. (1956). On the determination of probability distributions of more dimensions by their projections. Acta Mathematica Hungarica, 7(3):403410.

Hubel, D.H. and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1):106.

Hubel, D.H. and Wiesel, T.N. (1974). Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnifcation factor. The Journal of comparative neurology, (158):295-305.

Hung,C.P., Kreiman G., Poggio T., DiCarlo J.J. (2004) Fast Readout of Object Identity from Macaque Inferior Temporal Cortex Science, Vol. 310 no. 5749 pp. 863-866.

Hyvrinen A. and Oja E.(2000). Independent component analysis: algorithms and applications. Neural Networks, 13:411-430.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems, volume 25, pages 1106,1114, Lake Tahoe, CA.

Jones, J. P. and Palmer, L. A (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology, 58(6):1233-1258.

Kanwisher,N., Functional specificity in the human brain: a window into the functional architecture of the mind, Proceedings of the National Academy of Sciences, 107, 11163, 25.

Karhunen, J. (1994). Stability of Oja's PCA subspace rule. Neural Computation, 6:739-747.

Kouh, M. and Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. Neural computation, 6, 20, 1427–1451, MIT Press.

Ko, E. Y., Leibo, J. Z., and Poggio, T. (2011). A hierarchical model of perspective-invariant scene identification. In Society for Neuroscience (486.16/OO26)., Washington D.C.

Ku, S., Tolias, A., Logothetis, N., and Goense, J. (2011). fMRI of the Face Processing Network in the Ventral Temporal Lobe of Awake and Anesthetized Macaques. Neuron, 70(2):352-362.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541-551.

LeCun, Y. and Huang, F.J. and Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting, CVPR, 2, II-97.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech,

and time series The handbook of brain theory and neural networks, 255-258.

Le, Q. V., Monga, R. , Devin, M. , Corrado, G. , Chen, K. , Ranzato, M., Dean, J. and Ng, A. Y. (2011). Building high-level features using large scale unsupervised learning. CoRR.

Lee, T. and Soatto, S. (2012). Video-based descriptors for object recognition. Image and Vision Computing. (3,99):1645-50.

Leibo, J. Z., Mutch, J., and Poggio, T. (2011b). Why The Brain Separates Face Recognition From Object Recognition. In Advances in Neural Information Processing Systems (NIPS), Granada, Spain.

Leibo J. Z., Mutch J., and Poggio T.(2011a). 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.

Leibo, J.Z., Liao, Q., Anselmi, F., Poggio, T., (2014). The invariance hypothesis implies domain-specificregions in visual cortex. Cbmm memo 14. biorxiv.

Leibo JZ, Anselmi F, Mutch J, Ebihara AF, Freiwald W, Poggio T. (2013) Viewinvariance and mirror-symmetric tuning in a model of the macaque face-processing system (2013). Computational and Systems Neuroscience (I-54). Salt Lake City, UT.

Levi, D.M. (2008) Crowding-an essential bottleneck for object recognition: a mini-review. Vision Res. 48, 635-654.

Levi D.M., Carney T. (2011) The effect of flankers on three tasks in central, peripheral and amblyopic vision. Journal of Vision 11():10, 1-23.

Liao, Q, Leibo, J.Z., Poggio, T. (2013) Learning invariant representations and applications to face verification. Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, NV.

Logothetis, N. , Pauls, J. , Bulthoff, H. and Poggio, T.(1994). View-dependent object recognition by monkeys. Current Biology, 4(5):401-414.

Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Current Biology, 5(5):552-563.

Logothetis N. K., and Sheinberg D. L. (1996) Visual Object Recognition, Annual Review of Neuroscience, Vol. 19: 577-621.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. , Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., and Tootell, R. B. (1995). Object-

related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proceedings of the National Academy of Sciences, 92(18):8135-8139.

Mallat S. Group invariant scattering (2012). Communications on Pure and Applied Mathemathics 65(10):1331-1398.

Mallat S. (2008). A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way. ISBN-10: 0123743702.

Marr, E. H. D. and Poggio, T. (1980) Smallest channel in early human vision. JOSA, (70,7):868-870.

Marr, D. and Poggio, T. (1976). 1977 "From understanding computation to understanding neural circuitry", in Neuronal Mechanisms in Visual Perception Eds E Poppel, R Held, J E Dowling Neurosciences, Research Program Bulletin 15 470-488 Mazer, J. A., Vinje, W. E., McDermott, J., Schiller, P. H., and Gallant, J. L. (2002). Spatial frequency and orientation tuning dynamics in area v1. PNAS.

Mel, Bartlett, W. (1997). SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition, Neural Computation, 4, 777-804.

Mutch, J. and Lowe, D.G. (2006). Multiclass object recognition with sparse, localized features. Computer Vision and Pattern Recognition, 11-18.

Mutch J, Leibo JZ, Smale S, Rosasco L, Poggio T. (2010) Neurons that confuse mirror-symmetric object views. MIT-CSAIL-TR-2010-062, CBCL-295.

Niell C. and Stryker M. Highly selective receptive fields in mouse visual cortex. Journal of Neuroscience (2008) 28(30):7520-7536.

Niyogi, P., Poggio T., and Girosi F. (1998) Incorporating Prior Information in Machine Learning by Creating Virtual Examples. IEEE Proceedings on Intelligent Signal Processing.

Oja, E. (1982). Simplified neuron model as a principal component analyzer, Journal of mathematical biology, 15, 3, 267-273.

Oja E.(1992). Principal components, minor components, and linear neural networks. Neural Networks, 5(6):927-935.

Pelli, D.G. and Tillman, K.A. (2008) The uncrowded window of object recognition. Nat. Neurosci. 11, 1129-1135.

Perrett, D.I., and Oram, M.W. (1993), Neurophysiology of shape processing

Image and Vision Computing, 6, Vol 11, 317–333, Elsevier.

Pinto, N. and DiCarlo, James J. and Cox, D.D. (2009). How far can you get with a modern face recognition test set using only simple features?, Computer Vision and Pattern Recognition, 2591-2598.

Poggio, T., and Edelman, S. (1990). A network that learns to recognize three dimensional objects. Nature, 343(6255):263-266.

Poggio, T., J. Mutch, F. Anselmi, A. Tacchetti, L. Rosasco, and J.Z. Leibo (2013), "Does invariant recognition predict tuning of neurons in sensory cortex?", MIT-CSAIL-TR-2013-019, CBCL-313, Massachusetts Institute of Technology, Cambridge, MA.

Poggio, T.A., Mutch, J., Isik, L. (2014). Computational role of eccentricity dependent cortical magnification. CBMM Memo No. 017. arXiv:1406.1770v1 . CBMM Funded.

Poggio, T. and Reichardt W. 1973 Considerations on models of movement detection. Kybernetik, 13, 223-7.

Potter M.C. (1975) Meaning in visual search. Science, 187:565-566.

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

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

Ruderman D. (1994). The statistics of natural images. Network: Computation in Neural Systems, 5:517-548.

Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A. , Khosla A., Bernstein M., Berg A. C. and Fei-Fei L.. ImageNet (2014). Large Scale Visual Recognition Challenge. arXiv:1409.0575.

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

Saxe, A.M. and Bhand, M. and Mudur, R. and Suresh, B. and Ng, A. Y.. (2011) Unsupervised learning models of primary cortical receptive fields and receptive field plasticity, Advances in Neural Information Processing Systems 24, 1971-1979.

Schacter, D. L., and Addis D. R. (2009). On the nature of medial temporal lobe

contributions to the constructive simulation of future events. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1521):1245- 1253.

Schyns, P. G. and Gosselin, F. (2003) . Diagnostic use of scale information for componential and holistic recognition. In: Peterson, M.A., Rhodes, G. (Eds.), Perception of Faces, Objects, And Scenes, 120-145.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T.(2005). 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.

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

Sohl-Dickstein J., Wang C. M., Olshausen B. A. An Unsupervised Algorithm For Learning Lie Group Transformations, arXiv:1001.1027 2010.

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

Strasburger, M. J. H., Rentschler I. (2011) Peripheral vision and pattern recognition: A review. Journal of Vision, (11(5):13):1-82.

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

Stryker M.P. (1991) Temporal associations. News and Views. Nature 354, 108 - 109.

Tarr, M. J. (1995) Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. Psychon Bull Rev. 1995 Mar;2(1):55-82.

Tarr, M. J. and Gauthier, I.(2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nature Neuroscience, 3:764-770.

Thorpe S.J., Fize D., and Marlot C. (1996) Speed of processing in the human visual system. Nature, 381:520-522.

Torralba A. and Oliva A. (2003). Statistics of natural image categories. In Network: Computation in Neural Systems, pages 391-412.

Tsao, D. and Freiwald, W. (2003). Faces and objects in macaque cerebral cortex.

Nature, 6(9):989-995.

Tsao, D. and Livingstone M.S. (2008) Mechanisms of face perception Annu. Rev. Neurosci. 31: 411-437.

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

Ullman, S. and Basri, R. (1991) Recognition by linear combinations of models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (Volume:13 , Issue: 10), 992-1006.

Zeiler M.D., Fergus R. (2014), Visualizing and Understanding Convolutional Networks, ECCV.