Visual Cortex and Deep Networks

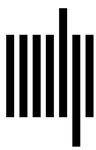
Learning Invariant Representations

By: Tomaso A. Poggio, Fabio Anselmi

Citation:

Visual Cortex and Deep Networks: Learning Invariant Representations
By: Tomaso A. Poggio, Fabio Anselmi
DOI: 10.7551/mitpress/10177.001.0001
ISBN (electronic): 9780262336710
Publisher: The MIT Press
Published: 2016

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

1 Invariant Representations

Mathematics of Invariance

1.1 Introduction and Motivation

One could argue that the most important aspect of intelligence is the ability to learn [16]. How do present supervised learning algorithms compare with brains? One of the most obvious differences is the ability of people and animals to learn from very few labeled examples. A child, or a monkey, can learn a recognition task from just a few examples. The main motivation of this book is the conjecture that the key to reducing the sample complexity of object recognition is invariance to transformations. Images of the same object usually differ from each other because of simple transformations such as translation, scale (distance), or more complex deformations such as viewpoint (rotation in depth) or change in pose (of a body) or expression (of a face).

The conjecture is supported by previous theoretical work showing that *almost all the complexity* in recognition tasks is due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the object [17]. It implies that in many cases recognition, both identification (e.g., of a specific car relative to other cars) and 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 like an individual face is only due to transformations. In the case of categorization, consider the suggestive evidence from the classification task in figure 1.1. 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 versus 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 (see box 1.1 and figure 1.1).

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

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

- The first 100 msec of vision in the ventral stream are mostly feedforward. The main computation goal is to generate a number of image representations, each one quasi-invariant to some transformations experienced during development, 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 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. Note that in a Turing test of vision [133] only the simplest, standard questions (what is there? who is there? etc.) can be answered by pretrained classifiers.

We consider only the feedforward architecture of the ventral stream and its computational function. To help readers understand more easily the mathematics of this part of the book, we anticipate here the network of visual areas that we propose for computing invariant representations for feedforward visual

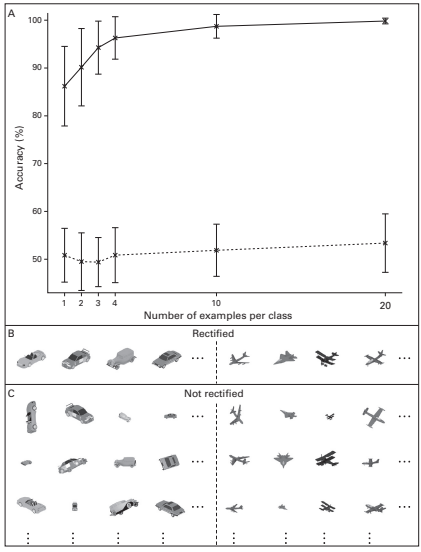


Figure 1.1

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

recognition. There are two main stages. The first one comprises retinotopic areas computing a representation that is invariant to affine transformations. The second computes approximate invariance to object-specific, nongroup transformations. The second stage consists of parallel pathways, each one for a different object class (see figure 1.5, stage 1). The results of this part do not strictly require these two stages. If both are present, as it seems is the case for the primate ventral stream, the mathematics of the theory requires that the object-specific stage follow the one dealing with affine transformations. According to the i-theory, the Hubel-Wised module is the basic module for both stages. The first- and second-stage pathways may each 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 in the mouse. The final architecture is shown in figure 1.5. 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. Note that for any representation that is invariant to X and selective for Y, there may be a dual representation that is invariant to Y but selective for X. In general, they are both needed for different tasks, and both can be computed by an HW module with different pooling strategies. In general, the circuits computing them share a good deal of overlap. For example, it is possible that different face patches in the cortex are used to represent different combinations of invariance and selectivity.

1.2 Invariance Reduces Sample Complexity of Learning

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: people do not have any problem in recognizing in a distance-invariant way a face seen only once. It is intuitive that representations of images that are invariant to transformations such as scaling, illumination, and pose should allow supervised learning from far fewer examples.

A proof of the conjecture for the special case of translation or scale or rotation is provided in appendix section A.1. For images defined on a grid of pixels, the result (in the case of translations) can be proved using well-known relations between covering numbers and sample complexity.

Box 1.1
Sample Complexity

Sample complexity is the number of examples needed for the estimate of a target function to be within a given error rate. In the example of figure1.1, the number of airplanes or cars, we trained the linear classifier to perform the recognition task with a certain precision.

Sample Complexity for Translation Invariance

Consider a space of images of dimensions $p \times p$, which may appear in any position within a window of size $rp \times rp$. The natural image representation yields a sample complexity (for a linear classifier) of order $m_{\text{image}} = O(r^2p^2)$; the invariant representation yields a sample complexity of order

$$m_{\text{inv}} = O(p^2). \tag{1.1}$$

This simple observation says that in the case of a translation group, 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 is that the sample complexity gain is on the order of the number of virtual examples generated by the action of the group on a single image (see [18, 19]). This is not a constructive result, but it supports the hypothesis that the ventral stream in the visual cortex tries to approximate such an oracle. The next section describes a biologically plausible algorithm that the ventral stream may use to implement an invariant representation.

1.3 Unsupervised Learning and Computation of an Invariant Signature (One-Layer Architecture)

The following algorithm is biologically plausible, as we discuss in detail in chapter 2, where we argue that it may be implemented in the cortex by an HW module, that is, a set of *KH* complex cells (see box 1.2) with the same receptive field, each pooling the output of a set of simple cells whose sets of synaptic weights correspond to 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, $h = 1, \cdots, H,\ \Delta > 0$.

Box 1.2

Simple Cells, Complex Cells [20]

A threshold vector product can be interpreted as the output of a neuron, called *simple cell*, which computes a possibly high-dimensional inner product with a template t and applies a nonlinear operation to it. In this interpretation, eq. (1.2) can be seen as the output of the pooling of many simple cells by a second neuron *complex cell*, which aggregates the output of other neurons by a simple averaging operation. Neurons of the former kind can be found in the visual cortex.

The algorithm for groups (finite or compact, defined on a finite or compact set) (see figure 1.2) is as follows.

Developmental stage

1. For each of K isolated (on an empty background) objects, or templates, memorize a sequence Λ of $|G|$ frames corresponding to the object's transformations $(g_i, i = 1, \cdots, |G|)$. For now, we suppose the g_i to belong to a finite group G; see box 1.3. The sequence of frames is observed over a time

interval; thus $\Lambda = \{g_0 t, g_1 t, \cdots, g_{|G|} t\}$ for template t. (For template t^k the corresponding sequence of transformations is denoted Λ_k.)

2. Repeat for each of K templates.

Runtime computation of invariant signature for a single image of any new object

1. For each t^k compute the dot product of the image with each of the $|G|$ transformations in Λ_k.
2. For each k compute the cumulative histogram of the resulting values.
3. The signature is the set of K cumulative histograms, that is, the set of

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\left\langle I, g_i t^k \right\rangle + h\Delta), \tag{1.2}$$

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

Box 1.3
Group Transformations

A group is a set of objects equipped with a law of composition, $*$, such that

- a composition of any two elements of the set gives another element of the set (closure);
- the composition is associative, i.e., $(a * b) * c = a * (b * c)$ (associativity);
- the set includes an identity element $a * e = a$ (identity);
- the set includes all the inverses of the elements, a^{-1} (inverse).

The algorithm consists of two parts. The first part is unsupervised learning of transformations by storing transformed templates, which are images. This part is possibly done only once during development of the visual system. The second part is the actual computation of invariant signatures during visual perception. Our analysis is not restricted to the case of group transformations. For now, we consider groups that are compact and, for simplicity, finite.

This algorithm we used throughout the book. The guarantees we can provide depend on the type of transformations. The main questions are whether the signature is invariant under the same type of transformations that were observed in the first stage, and whether the signature is selective, for instance, can it distinguish between N different objects. In summary, as shown in appendix section A.2, the HW algorithm is invariant and selective (when K

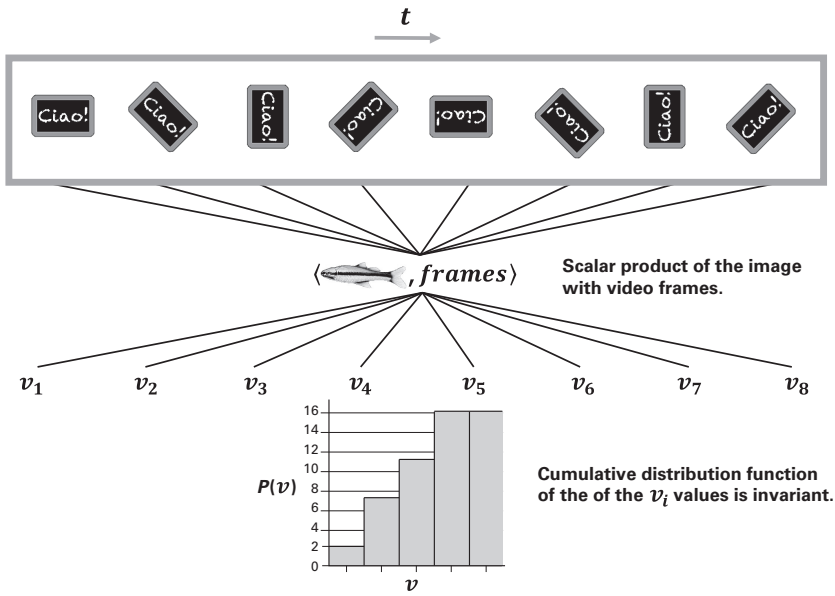


Figure 1.2

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 (see eq. (1.2)) corresponds to the the histogram where, e.g., $k = 1$ denotes the template blackboard, h the bins of the histogram; transformations g_i 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 by storing sets of transformed templates that are unrelated to the object to be represented in an invariant way. In particular, the templates could be random patterns.

is order $\log(N)$) if the transformations form a group (but see also appendix section A.2.4). In this case, any set of randomly chosen templates will work for the first stage. Seen as transformations from a 2-D image to a 2-D image, the natural choice is the affine group consisting of translations, rotations in the image plane, scaling (possibly nonisotropic), and compositions thereof. The HW algorithm can learn with exact invariance and desired selectivity in the case of the affine group or its subgroups. In the case of 3-D images consisting of voxels with x, y, z coordinates, rotations in 3-D 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). In section 1.6 we show that the same HW algorithm provides approximate invariance (under some conditions) for non-group transformations such as the transformations from 3-D to 2-D induced by 3-D 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 appendix section A.2 and in [8, 10]).

Box 1.4
Invariance and Group Average

For any function $f \in L^2(\mathbb{R})$ and compact group G,

$$\bar{f}(x) = \int dg\, f(gx)$$

is invariant. This property can be simply proved using the fact that the group is closed under its composition rule and the invariance of the Haar measure dg:

$$\bar{f}(\tilde{g}x) = \int dg\, f(\tilde{g}gx) = \int dg\, f(gx) = \bar{f}(x), \quad \forall \tilde{g} \in G,$$

i.e., $\bar{f}$ is invariant to G.

Invariance Theorem

The distributions represented by eq. (1.2) are invariant, that is, each bin is invariant

$$\mu_h^k(I) = \mu_h^k(gI) \tag{1.3}$$

for any g in G, where G is the group of transformations labeled g_i in eq. (1.2). The proof is based on the fact that the signature is a group average (see box 1.4).

Selectivity Theorem

For groups of transformations (e.g., the affine group), the distributions represented by eq. (1.2) 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

$$K > \frac{c}{\epsilon^2} \ln \frac{N}{\delta}, \tag{1.4}$$

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. The stability of the signature under noisy measurements remains an open problem. Because

of the restricted dynamic range of cortical cells, the number H of bins is likely to be small, probably around 2 or 3 [21]. Note that other, related representations are possible (see [22]). 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 [23]) or the max,

$$\mu^k_{\text{av}}(I) = \frac{1}{|G|}\sum_{i=1}^{|G|}\left\langle I, g_i t^k\right\rangle,$$

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

$$\mu^k_{\text{max}} = \max_{g_i \in G}\left\langle I, g_i t^k\right\rangle,$$

can in practice replace the cumulative distribution function. Any linear combination of the moments is also invariant, and a small number of linear combinations is likely to be sufficiently selective.

1.4 Partially Observable Groups

This section outlines invariance, uniqueness, and stability properties of the signature obtained in the case in which transformations of a group are observable only within a *window* on the orbit (figure 1.3). The term *partially observable groups* (POGs) emphasizes the properties of the group—in particular, associated invariants—as seen by an observer (e.g., a neuron) looking through a window at a part of the orbit. Therefore the window should not be thought of only as a spatial window but over ranges of transformation parameters, for example, a window over scale and spatial transformations. With this observation in mind, let G be a finite group and $G_0 \subseteq G$ a subset (G_0 is not usually a subgroup). The subset of transformations G_0 can be seen as the set of transformations that can be *observed* through a window on the orbit (i.e., the transformations that correspond to a part of the orbit). A *local* signature associated with the partial observation of G can be defined considering

$$\mu^k_h(I) = \frac{1}{|G_0|}\sum_{i=1}^{|G_0|}\eta_h\left(\left\langle I, g_i t^k\right\rangle\right) \tag{1.5}$$

where η_h is a set of nonlinear functions and $\Sigma_{G_0}(I) = (\mu^k_h(I))_{h,k}$. This definition can be generalized to any locally compact group considering

$$\mu^k_h(I) = \frac{1}{V_0}\int_{G_0}\eta_h\left(\left\langle I, gt^k\right\rangle\right)dg, \quad V_0 = \int_{G_0} dg. \tag{1.6}$$

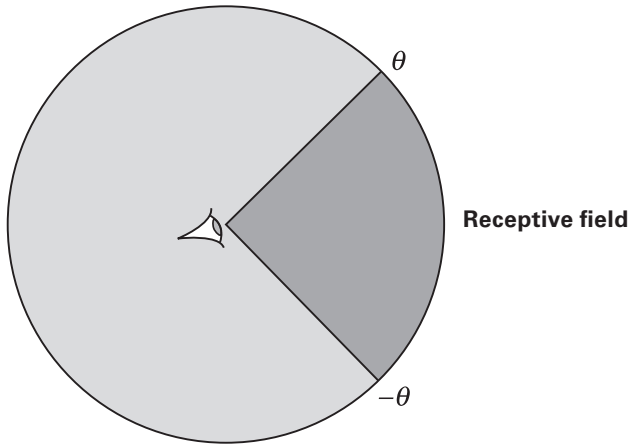


Figure 1.3

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

The constant V_0 normalizes the Haar measure dg, restricted to G_0, so that it defines a probability distribution. The latter is the distribution of the images subject to the group transformations that are observable, that is, in G_0. These definitions can be compared to the definition in eq. (1.2) in the fully observable group case. We next discuss the properties of this signature. While uniqueness follows essentially from the analysis so far, invariance requires a new analysis.

1.5 Optimal Templates for Scale and Position Invariance Are Gabor Functions

The previous results apply to all groups, in particular to those that 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 (see appendix section A.2.5 but also [1, 8]). For the standard 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 proved (see appendix section A.2.5) that the templates must be maximally localized relative to generic input images. In the case of translation and scale invariance, this requirement is equivalent to localization in 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 [24, 25] that they must have a Gaussian envelope (for a certain definition of minimum spatial and frequency size). If the system of templates is required to be a frame then the following property holds.

Optimal Invariance Theorem

Gabor functions (here in $1-D$) of the form $t(x) = e^{-x^2/(2\sigma^2)}e^{i\omega_0 x}$ are the templates that give simultaneous maximal invariance 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 (for a specific definition of *optimal*). This is the only computational justification we know of for the Gabor shape of simple cells in V1, which seems to be remarkably universal: it holds in primates [26], cats [27], and mice [28] (see figure 1.4 for results of simulations).

1.6 Quasi Invariance to Nongroup Transformations Requires Class-Specific Templates

All the previous results require a group structure and ensure exact invariance for a single new image. In 2-D all combinations of translation, scaling, and rotation in the image plane are included; transformations induced on the image plane by 3-D transformations, such as viewpoint changes and rotation in depth of an object, are not. The latter form a group in 3-D, that is, if images and templates were 3-D views. In principle, motion or stereopsis can provide the third dimensional though available psychophysical evidence [29, 30] suggests that human vision does not use it for recognition. Note that transformations in the image plane are affected not only by orthographic projection of the 3-D geometry but also by the process of image formation, which depends on the 3-D geometry of the object, its reflectance properties, and the relative location of light source and viewer.

It turns out that the same HW algorithm can still be applied to nongroup transformations such as transformations of the expression of a face or 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

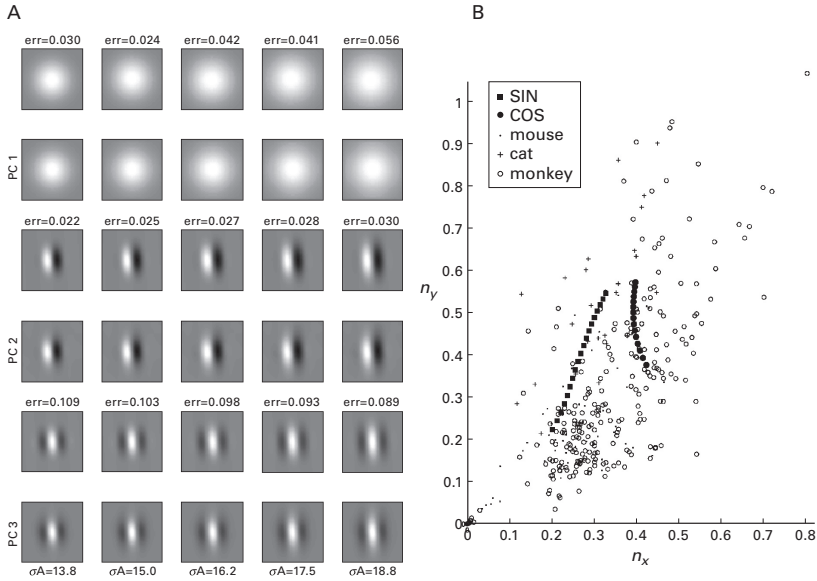


Figure 1.4

(A) Simulation results for V1 simple cells learning via principal component analysis. 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) $n_y = \sigma_y/\lambda$ vs $n_x = \sigma_x/\lambda$ for the modulated (x) and unmodulated (y) direction of the Gabor wavelet. 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. Source: [11].

results. The key technical requirement is that a condition of incoherence (see box 1.5) with respect to a transformation hold for the class of images I_C with respect to the dictionary t^k under the transformation T_r (we consider here a one-parameter transformation, r, for simplicity),

Box 1.5

Sparsity/Incoherence

A signal is sparse if its information content can be expressed with few coefficients when expanded in the right basis. For example, a sinusoid in the Fourier space is expressed by one coefficient, and therefore its representation is very sparse in the Fourier basis.

Two bases are called incoherent when signals are sparse when expressed in one basis and dense in the other. For a sinusoidal signal, the Fourier and time domain are incoherent. Incoherence extends the space-frequency duality to other bases.

$$\left\langle I_C, T_r t^k \right\rangle \approx 0, \quad |r| > a, \;\; a > 0. \tag{1.7}$$

Strictly speaking, the condition is valid when the object has localized support in the pooling region (is an isolated object). However, it holds approximately whenever the dot product has a fast decay with the transformation (e.g., wavelet coefficients). The property, which is an extension of the compressive sensing notion of mutual coherence, requires that the templates (an image can be considered a template) have a representation with sharply peaked correlation and autocorrelation (the constant a above in eq. (1.7) is related to the support of the peak of the correlation). Eq. (1.7) can be satisfied by templates that are similar to images in the set and are sufficiently rich to be incoherent for small transformations. Empirically, it appears, our incoherence condition is usually satisfied by the neural representation of images and templates at some high level of the hierarchy of HW modules. Like standard mutual incoherence (see [25]) our condition of incoherence with respect to a group is generic. Most neural patterns (templates and images from the same class) chosen at random will satisfy it. The full (theorem A.8 in appendix section A.2.5) takes the following form.

Class-Specific Property

$\mu_h^k(I)$ is approximatively invariant around a view if

- the dictionary of the templates relative to the transformations is incoherent,
- I is one of the templates and transforms in the same way as the templates;
- the transformation is smooth.

The main implication is that approximate invariance can be obtained for nongroup 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. An example is face-tuned cells in the face patches. Unlike exact invariance for affine transformations, where tuning of the simple cells is nonspecific in the sense that does not depend on the type of image, nongroup transformations require highly tuned neurons and yield at best only approximate invariance.

Summary of the Results

The core of i-theory applies without qualification to compact groups such as rotations of the image in the image plane. Translation and scaling are, however, only locally compact. Each HW module usually observes only a part of the transformation's full range. Each module has a finite pooling range, corresponding to a finite window on the orbit associated with an image.

Exact invariance for each module is equivalent to a condition of *localization/sparsity* of the dot product between image and template. In the simple case of a group parameterized by one parameter r the condition is

$$\left\langle I, g_r t^k \right\rangle = 0 \quad |r| > a. \tag{1.8}$$

Since this condition is a form of sparsity of the generic image I with respect to a dictionary of templates t^k (under a group), this result may provide a justification for *sparse* encoding in sensory cortex (see, e.g., [31]). Localization yields the following surprising result: *optimal invariance for translation and scale implies Gabor functions as templates*. Since a frame of Gabor wavelets follows from natural requirements of completeness, this may also provide a general motivation for the scattering transform approach of Mallat based on wavelets [32]. Eq. (1.8), if relaxed to hold approximately, that is $\left\langle I_C, g_r t^k \right\rangle \approx 0 \quad |r| > a$, becomes a *sparsity condition for the class of I_C with respect to the dictionary t^k under the generic transformation T* (approximated locally by the group transformation g_r) when restricted to a subclass I_C of similar images. This property, which is similar to compressive sensing incoherence (but in a group context), requires that I and t^k have a representation with rather sharply peaked autocorrelation (and correlation). When the condition is satisfied, the basic module equipped with such templates can provide *approximate invariance* to nongroup transformations, such as rotations in depth of a face or its changes of expression (see appendix section A.2 and A.8). In summary, condition Eq. 1.8 can be satisfied in two different *regimes*. The first one, exact and valid for generic I, yields optimal Gabor templates. The second regime, approximate and valid for specific subclasses of I, yields highly tuned templates, specific for the subclass. This argument suggests generic, Gabor-like templates in the first layers of the hierarchy and highly specific templates at higher levels (incoherence improves with increasing dimensionality; see appendix section A.2.5).

1.7 Two Stages in the Computation of an Invariant Signature: Extension of the HW Module to Hierarchical Architectures

It is known that Hubel and Wiesel's original proposal [20] for visual area V1—of a module consisting of complex cells (C-units) combining the outputs of sets of simple cells (S-units) with identical orientation preferences but differing retinal positions—can be used to construct group-invariant detectors. This is the insight underlying many networks for visual recognition, including HMAX [33] and convolutional neural nets [2, 34]. We showed that a representation of images and image patches, in terms of a feature vector that is invariant to a broad range of transformations, such as translation, scale, expression of a

face, pose of a body, and viewpoint, makes it possible to recognize objects from only a few labeled examples. In the following we argue that hierarchical architectures of HW modules (indicated by $\bigwedge$ in figure 1.5) can provide such invariant representations while maintaining discriminative information about the original image. Each $\bigwedge$-module provides a feature vector (the signature) for the the part of the visual field that is inside its receptive field; the signature is invariant to ($\mathbb{R}^2$) affine transformations within the receptive field. The hierarchical architecture, since it computes a set of signatures for different parts of the image, is invariant to the more general family of locally affine transformations (which include globally affine transformations of the whole image). This reasoning also applies to invariance to global transformations that are not affine but are locally affine, that is, affine within the pooling range of some of the modules in the hierarchy (any differentiable transformation, no matter how complex, can be seen locally as an affine transformation). This remarkable invariance of the hierarchies follows from the key property of *covariance* (see box 1.6) of such architectures for certain image transformations and from the uniqueness and invariance of the individual module signatures. The basic HW module (see section 1.3) is at the core of the properties of the architecture.

Box 1.6
Covariance

In our theory, the key property of hierarchical architectures of layers of repeated HW modules—allowing the recursive use of single module properties at all layers—is the property of *covariance*: the response image at layer ℓ transforms like the response image at layer $\ell - 1$.

The main reasons for an extension to a hierarchical architecture like the are shown in figure 1.5 are the following:

- *Compositionality*. A hierarchical architecture provides signatures of larger and larger patches of the image in terms of lower-level signatures. Because of this, it can access memory in a way that matches naturally with the linguistic ability to describe a scene as a whole and as a hierarchy of parts.
- *Approximate factorization*. In architectures such as the network sketched in figure 1.5, approximate invariance to transformations specific for an object class can be learned and computed in different stages. This property may provide an advantage in terms of the sample complexity of multistage learning [16]. For instance, approximate class-specific invariance to pose (e.g., for faces) can be computed on top of a translation-and-scale-invariant

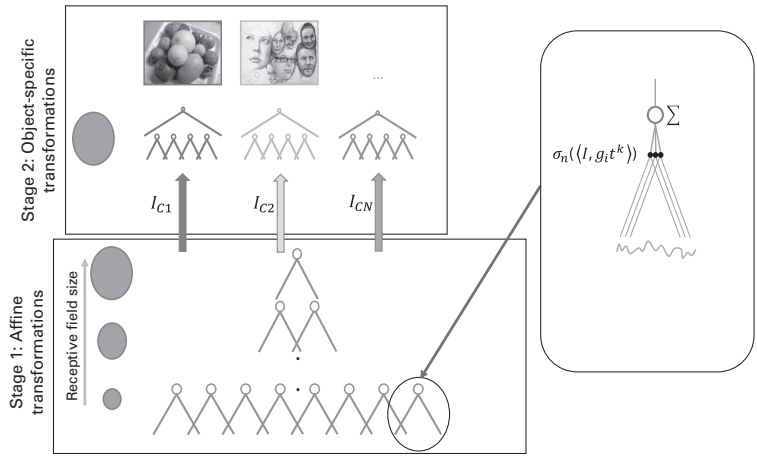


Figure 1.5

A hierarchical architecture of HW modules, indicated by $\bigwedge$. Signature provided by each of the nodes at each layer may be used by a supervised classifier. *Stage 1:* a hierarchy of HW modules *(inset)* with growing receptive fields provides a final signature (top of the hierarchy) that 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 nongroup transformations (e.g., rotation in depth).

representation [35]. Thus the implementation of invariance can, in some cases, be factorized into different steps corresponding to different transformations (see [36, 37] for related ideas).

- *Optimization of local connections* and optimal reuse of computational elements. Despite the high number of synapses on each neuron it would be impossible for a complex cell to pool information across all the simple cells needed to cover an entire image.
- *Minimization of number of transformations* per template needed to achieve invariance (see appendix section A.10).
- *Clutter tolerance* (see chapter 3).

Probably all these properties together are the reason that evolution developed hierarchies.

One-layer architectures are unable to capture the *hierarchical organization of the visual world*, where scenes are composed of objects that are themselves composed of parts. Objects (parts) can change position or scale in a scene relative to each other without changing their identity and often changing the scene only in a minor way. Thus global and local signatures from all levels of the hierarchy must be able to access memory in order to enable the categorization

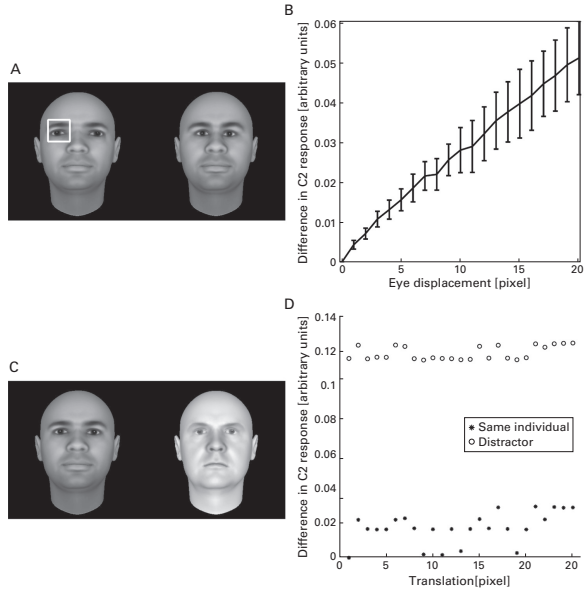


Figure 1.6

Empirical demonstration of the properties of invariance, stability, and uniqueness of the hierarchical architecture in a specific two-layer implementation (HMAX). (A) reference image on the left and a deformation of it (the eyes are closer to each other) on the right. (B) HW module at layer 2 (c_2) whose receptive fields contain the whole face provides a signature vector that is (Lipschitz) stable with respect to the deformation. In all cases, the figure shows just the Euclidean norm of the signature vector. The c_1 and c_2 vectors are not only invariant but also selective. Error bars represent ± 1 standard deviation. (C) two different images are presented at various locations in the visual field. (D) Euclidean distance between the signatures of a set of HW modules at layer 2 with the same receptive field (the whole image) and a reference vector. The signature vector is invariant to global translation and discriminative (between the two faces). In this example the HW module represents the top of a hierarchical, convolutional architecture. The images used were 200×200 pixels.

and identification of whole scenes as well as patches of the image corresponding to objects and their parts. Figure 1.6 shows examples of invariance and stability for wholes and parts. In the architecture of figure 1.5, each $\bigwedge$-module provides uniqueness, invariance, and stability at different levels, over increasing ranges from bottom to top. Thus, in addition to the desired properties of invariance, stability, and discriminability, these architectures match the hierarchical structure of the visual world and the need to retrieve items from memory at various levels of size and complexity.

The property of *compositionality* is related to the efficacy of hierarchical architectures versus one-layer architectures in dealing with the problem of

partial occlusion and the more difficult problem of clutter in object recognition. Hierarchical architectures are better at recognition in clutter than one-layer networks (see chapter 3 and [38]) because they provide signatures for image patches of several sizes and locations. However, hierarchical feedforward architectures cannot fully solve the problem of clutter. More complex (e.g., recurrent) architectures are likely needed for human-level recognition in clutter (see, e.g., [39–41]) and for other aspects of human vision.

A hierarchical architecture has layers with receptive fields of increasing size (figure 1.7). The intuition is that transformations represented at each level of the hierarchy begin with small affine transformations that is, over a small range of translation, scale, and rotation. The size of the transformations represented in the set of transformed templates will increase with the level of the hierarchy and the size of the apertures. In addition it seems intuitive that only translations will be seen by small apertures with scale and orientation changes being relevant later in the hierarchy.

To be more specific, suppose that the first layer consists of an array of small apertures (in fact corresponding to the receptive fields of V1 cells) and focus on one of the apertures. Box 1.7 explains why the only transformations that can

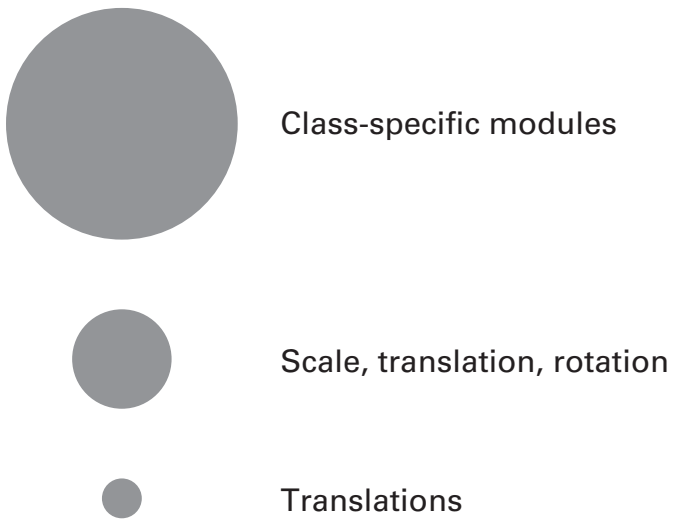


Figure 1.7

The conjecture is that receptive field sizes (of complex cells) affect not only the size but also the type of transformations that are learned and represented by the templates. In particular, small apertures (such as in complex cells in V1) only see small translations.

be seen by a small aperture are small translations, even if the transformation of the image is more complex.

Box 1.7

Any transformation looked at through a small window is approximately an affine transformation.

Our approach differs in the assumption that small (close to identity) diffeomorphic transformations can be well approximated at the first layer as locally affine transformations or, in the limit, as local translations, which therefore falls into the partially observable groups case (see section 1.4). This assumption is substantiated by the following reasoning (here in 1-D for simplicity), in which any smooth transformation is seen as parameterized by the parameter r and expanded around zero as

$$T_r(I) = T_0(I) + A_r(I) + O(r) = (e + A_r)(I) + O(r),$$

where $e + A_r$ is a linear operator in $GL(\mathbb{R})$, the general linear group.

In our theory, the key property of hierarchical architectures of repeated HW modules—allowing the recursive use of single module properties at all layers—is the property of *covariance*.

It is illuminating to consider two extreme cartoon architectures for the first of the two stages described in figure 1.5:

- One layer comprising one HW module and its *KH* complex cells, each 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 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 nongroup transformations directly into this one hidden layer architecture.

However, an HW module dealing with nongroup transformations can be added on top of the affine module. The theorems allow for this factorization (see appendix section A.9). 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, the mathematics allows for factorization of the range of invariance for the same group of transformations. This justifies the first layers of the second architecture, corresponding to figure 1.5, stage 1, where the size of the receptive field of each HW module and the range of its invariance increases, from lower to higher layers.

One of the main problems with the one-layer architecture is that it provides an invariant signature to an isolated object but not to its parts. This problem of recognizing wholes and parts is closely related to the problem of recognizing objects in clutter, recognizing an object independently of the presence of another one nearby. 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. 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 appendix section A.3). All these signatures could then be used by class-specific modules, possibly in a reverse hierarchy strategy (see [42]), that is, using first the top-level signatures and then the low-level ones in a task-dependent, top-down mode. We describe this architecture, starting with the retina and V1, in chapter 3. Three points are of interest here:

- Factorization of the 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, for instance, 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 according to the same group of transformations. It turns out that the architectures we describe have this property (see box 1.7): isotropic architectures, like the ones considered in this book, with pointwise 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 increases from one layer to the next, there is an invariance over larger and larger transformations. 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) but from the other levels as well (directly or indirectly).
- In the case of a discrete group one can prove that the number of different templates K required for selectivity is $K = C(\epsilon)(\log(n|G|)$ (see eq. (1.4)); that means it may significantly depend on the size of the group G that is pooled (see appendix section A.2.4).

1.8 Deep Networks and i-Theory

The class of learning algorithms called deep learning (particularly convolutional networks) is based on two computational operations in multiple layers.

The first operation is the inner product of an input with another point called a template (or filter or kernel), followed by a nonlinearity. The output of the dot product corresponds to the neural response of a simple cell [20]:

$$I \mapsto |\langle I, gt\rangle + b|_+,$$

where $|y|_+ = \max,(0, y)$.

The second operation is *pooling*. It aggregates in a single output the values of the different inner products computed, for example, via a sum

$$\sum_g |\langle I, gt\rangle + b|_+, \quad t \in \mathcal{T}, b \in \mathbb{R} \tag{1.9}$$

or a max operation $\max_g |\langle I, gt\rangle + b|_+$. This corresponds to the neural response of a complex cell [20]. In i-theory the nonlinearity acting on the dot product is considered part of the computation of a histogram, but the networks are completely equivalent.

It turns out (see references in [12]) that units of a deep convolutional network (DCN) using linear rectifiers (called by Breiman "ramps") correspond to a kernel with

$$\tilde{K}(I, I') = \int dg\, dg' \int dt\, db\, |\langle gt, I\rangle + b|_+ |\langle g't, I'\rangle + b|_+. \tag{1.10}$$

This result shows that linear rectifiers (and other nonlinear operations) in a deep network are equivalent to replacing the plain dot product with dot products in the feature space defined by the kernel. All the results described in the book hold for networks with a broad range of nonlinearities after the dot product.

Further extension are possible. Ongoing work [46] shows that i-theory can be formulated to include present-day deep convolutional learning networks (DCLNs) with supervision and nonpooling layers. In particular [46] shows how defining an extension of classical additive splines for multivariate function approximation (called hierarchical splines) is possible to have a theoretical framework for DCLNs with linear rectifers and pooling (sum or max).

The comparison of hierarchical versus shallow architecture in the context of i-theory and its extensions is explored in [47]. Hierarchical as well as shallow

networks can approximate functions of several variables, in particular those that are compositions of low-dimensional functions. [47] provides a characterization of the power of deep network architectures with respect to shallow networks: in particular it proves how shallow networks can approximate compositional functions with the same error of a deep network at the cost of a VC-dimension (Vapnik-Chervonkis dimension, a measure of the network capacity) that is exponential rather than quadratic in the dimensionality of the function. Compositional functions are also shown to be critical for image recognition, thus demonstrating a theoretical reason why deep architectures outperform shallow ones in image recognition.

Historical Background and Bibliography

There exists an extensive literature about invariant approaches to image representation. One of the first and most inspiring papers that used explicitly the group structure of the transformations and the group average technique to build invariant features is Mirbach's [43]. A different approach using group-invariant kernels was analyzed in Burkhardt [44]. Together with the work on kernel average embedding [45], it inspired our recent work on the equivalence between deep convolutional networks and hierarchical kernel machines [12]. More recently Mallat [32] developed an invariant representation called group scattering, which uses a cascade of modulus nonlinearities applied to wavelets coefficients in order to get group-invariant representations that are robust to small diffeomorphic transformations.

Chapter 1 has described work in our group. Its main sources are the following technical reports and journal papers: [1, 8, 9, 11, 12].

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.

Library of Congress Cataloging-in-Publication Data

Names: Poggio, Tomaso, author. | Anselmi, Fabio, author.
Title: Visual cortex and deep networks : learning invariant representations / Tomaso A. Poggio and Fabio Anselmi.
Description: Cambridge, MA : MIT Press, [2016] | Series: Computational neuroscience | Includes bibliographical references and index.
Identifiers: LCCN 2016005774 | ISBN 9780262034722 (hardcover : alk. paper)
Subjects: LCSH: Visual cortex. | Vision. | Neural networks (Neurobiology) | Perceptual learning. | Computational neuroscience.
Classification: LCC QP383.15 .P64 2016 | DDC 612.8—dc23 LC record available at http://lccn.loc.gov/2016005774

10 9 8 7 6 5 4 3 2 1