Unsupervised Learning of Invariant Representations in Hierarchical Architectures

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

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

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

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

Invariance | Hierarchy | Convolutional networks | Visual cortex

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

It is known that Hubel and Wiesel's original proposal [31] for visual area V1—of a module consisting of complex cells (C-units) combining the outputs of sets of simple cells (S-units) with identical orientation preferences but differing retinal positions—can be used to construct translation-invariant detectors. This is the insight underlying many networks for visual recognition, including HMAX [32] and convolutional neural nets [33, 34]. We show here how the original idea can be expanded into a comprehensive theory of visual recognition relevant for computer vision and possibly for visual cortex. The first step in the theory is the conjecture that a representation of images and image patches, with a feature vector that is invariant to a broad range of transformations—such as translation, scale, expression of a face, pose of a body, and viewpoint—makes it possible to recognize objects from only a few labeled labeled examples, as humans do. The second step is proving that hierarchical architectures of Hubel-Wiesel ('HW') modules (indicated by $\bigwedge$ in Fig. 1) can provide such invariant representations while maintaining discriminative information about the original image. Each $\bigwedge$-module provides a feature vector, which we call a *signature*, for the the part of the visual field that is inside its "receptive field"; the signature is invariant to ($\mathbb{R}^2$) affine transformations within the receptive field. The hierarchical architecture, since it computes a set of signatures for different parts of the image, is invariant to the rather general family of locally affine transformations (which includes globally affine transformations of the whole image). This remarkable invariance of the hierar-

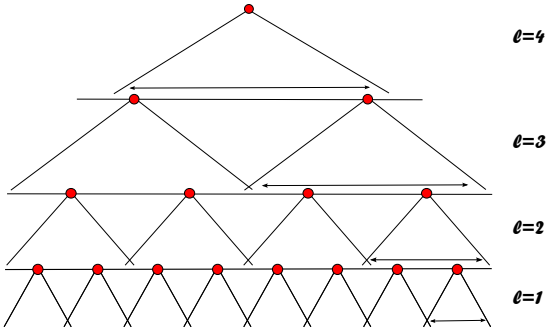

Fig. 1: *A hierarchical architecture built from HW-modules. Each red circle represents the signature vector computed by the associated module (the outputs of complex cells) and double arrows represent its receptive fields – the part of the (neural) image visible to the module (for translations this is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

Reserved for Publication Footnotes

[1]At the time of our writing, the working monograph [35] contains the most up-to-date account of the theory. The current monograph evolved from one that first appeared in July 2011 ([35]). Shorter papers describing isolated aspects of the theory have also appeared: [36, 37, 35]. The present paper is the first time the entire argument has been brought together in a short document.

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

Invariant representations and sample complexity

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

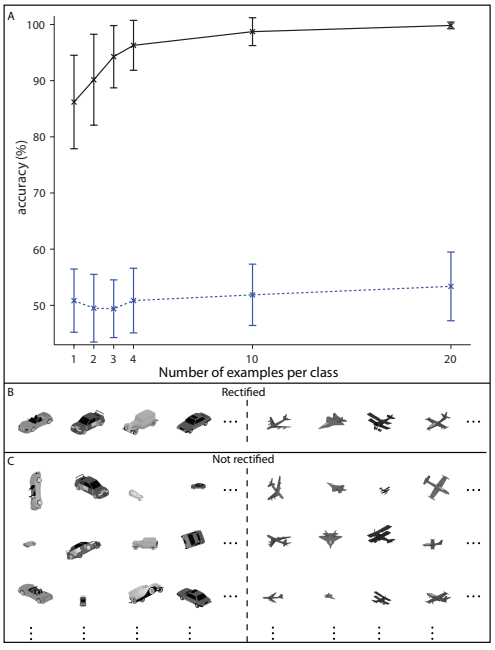

Fig. 2: *Sample complexity for the task of categorizing cars vs airplanes from their raw pixel representations (no preprocessing). A. Performance of a nearest-neighbor classifier (distance metric = 1 - correlation) as a function of the number of examples per class used for training. Each test used 74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

The conjecture is supported by previous theoretical work showing that *almost all the complexity* in recognition tasks is often due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the object [38]. It implies that in many cases, recognition—i.e., both identification, e.g., of a specific car relative to other cars—as well as categorization, e.g., distinguishing between cars and airplanes—would be much easier (only a small number of training examples would be needed to achieve a given level of performance), *if* the images of objects were rectified with respect to all transformations, or equivalently, if the image representation itself were invariant.

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

Invariance and uniqueness

Consider the problem of recognizing an image, or an image patch, independently of whether it has been transformed by the action of a group like the affine group in $\mathbb{R}^2$. We would like to associate to each object/image I a *signature*, i.e. a vector which is *unique* and *invariant* with respect to a group of transformations, but our analysis, as we will see later, is not restricted to the case of groups. In the following, we will consider groups that are compact and, for simplicity, finite (of cardinality $|G|$). We indicate, with slight abuse of notation, a generic group element and its (unitary) representation with the same symbol g, and its action on an image as $gI(x) = I(g^{-1}x)$ (e.g. a translation, $g_\xi I(x) = I(x - \xi)$). A natural mathematical object to consider is the *orbit* O_I—i.e., the set of images gI generated from a single image I under the action of the group. We say that two images are equivalent when they belong to the same orbit: $I \sim I'$ if $\exists g \in G$ such that $I' = gI$. This equivalence relation formalizes the idea that an orbit is invariant and unique. Indeed, if two orbits have a point in common they are identical everywhere. Conversely, two orbits are different if none of the images in one orbit coincide with any image in the other (see also [39]).

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

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

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

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

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

As argued later, simple operations for neurons are (high-dimensional) inner products, $\langle \cdot, \cdot \rangle$, between inputs and stored "templates" which are neural images. It turns out that classical results (such as the Cramer-Wold theorem [40], see Theorem 2 section 1 in SI Appendix) ensure that a probability distribution P_I can be almost uniquely characterized by K one-dimensional probability distributions $P_{\langle I,t^k \rangle}$ induced by the (one-dimensional) results of projections $\langle I, t^k \rangle$, where $t^k,\ k = 1, ..., K$ are a set of randomly chosen images called templates. A probability function in d variables (the image dimensionality) induces a unique set of 1-D projections which is discriminative; empirically a small number of projections is usually sufficient to discriminate among a finite number of different probability distributions. Theorem 3 in SI Appendix section 1 says (informally) that an approximately invariant and unique signature of an image I can be obtained from the estimates of K 1-D probability distributions $P_{\langle I,t^k \rangle}$ for $k = 1, \cdots, K$. The number K of projections needed to discriminate n orbits, induced by n images, up to precision ϵ (and with confidence $1 - \delta^2$) is $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant.

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

Memory-based learning of invariance

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

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

A theory of pooling

The arguments above make a few predictions. They require an effective normalization of the elements of the inner product (e.g. $\langle I, g_i t^k \rangle \mapsto \frac{\langle I, g_i t^k \rangle}{\|I\| \|g_i t^k\|}$) for the property $\langle gI, t^k \rangle = \langle I, g^{-1} t^k \rangle$ to be valid (see section 0 of SI Appendix). Notice that invariant signatures can be computed in several ways from one-dimensional probability distributions. Instead of the $\mu_n^k(I)$ components representing directly the empirical distribution, they may represent the moments $m_n^k(I) = 1/|G| \sum_{i=1}^{|G|} (\langle I, g_i t^k \rangle)^n$ of the same distribution [41]. Under weak conditions, the set of *all* moments uniquely characterizes the one-dimensional distribution $P_{\langle I,t^k \rangle}$ (and thus P_I). $n = 1$ corresponds to pooling via sum/average (and is the only pooling function that does not require a nonlinearity); $n = 2$ corresponds to "energy models" of complex cells and $n = \infty$ is related to the max-pooling. In our simulations, using just one of these moments seems to usually provide sufficient selectivity to a hierarchical architecture (see SI Appendix section 5). Other nonlinearities are also possible; see [35]. The arguments of this section may begin to provide a theoretical understanding of "pooling", giving insight to the search for the "best" choice in any particular setting—something which is normally done empirically for each application (e.g., [42]). According to this theory, these different pooling functions are all invariant, each one capturing part of the full information contained in the PDFs.

Implementations

There are other interesting and surprising results beyond the core of the theory described above. We sketch some of the main ones – the supplementary information provides the mathematical statements. Here it is important to stress that the theory has strong empirical support from several specific implementations which have been shown to perform well on a number of databases of natural images. The main set of tests is provided by HMAX, an architecture in which pooling is done with a max operation and invariance, to translation and scale, is mostly hardwired (instead of learned). Its performance on a variety of tasks is summarized in SI Appendix section 5. Strong performance is also achieved by other very similar architectures (again special cases of the theory) such as [43]. High performance for non-affine and even non-group transformations allowed by the hierarchical extension of the theory (see below) has been shown on large databases of face images, where our latest system advances the state-of-the-art on several tests [37]. Deep learning convolutional networks are another case of architectures that have achieved very good performance and are probably special cases of the theory even if they do not incorporate all of the possible invariances or their unsupervised learning ([44, 45], but see [46]).

Extensions of the Theory

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

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

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

Since this condition is a form of sparsity of the generic image I w.r.t. to a dictionary of templates t^k (under a group), this results provides a powerful justification for *sparse* encoding in sensory cortex (e.g. [47]).

It turns out that localization yields the following surprising result (Theorem 6 and 7 in SI Appendix): *optimal invariance for translation and scale implies Gabor functions as templates*. Since a frame of Gabor wavelets follows from natural requirements of completeness, this may also provide a general motivation for the Scattering Transform approach of Mallat based on wavelets [48].

The same Equation 15, if relaxed to hold approximately, that is $\left\langle I_C, g_r t^k \right\rangle \approx 0 \quad |r| > a$, becomes a *sparsity condition for the class of I_C wrt the dictionary t^k under the group G* when restricted to a subclass I_C of similar images. This property (see SI Appendix at the end of section 1), which is similar to compressive sensing "incoherence" (but in a group context), requires that I and t^k have a representation with rather sharply peaked autocorrelation (and correlation). When the condition is satisfied, the basic HW-module equipped with such templates can provide approximative invariance to non-group transformations such as rotations in depth of a face or its changes of expression (see Proposition 8, section 1, SI Appendix). In summary, condition Equation 15 can be satisfied in two different *regimes*. The first one, exact and valid for generic I, yields optimal Gabor templates. The second regime, approximate and valid for specific subclasses of I, yields highly tuned templates, specific for the subclass. Note that this arguments suggests generic, Gabor-like templates in the first layers of the hierarchy and highly specific templates at higher levels (note also that incoherence improves with increasing dimensionality).

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

One-layer networks can achieve invariance to *global* transformations of the whole image (exact invariance if the transformations are a subgroup of the affine group in $\mathbb{R}^2$) while providing a unique global signature which is stable with respect to small perturbations of the image, (see Theorem 4 SI Appendix and [35]). The two main reasons for a hierarchical architecture such as Fig. 1 are a) the need to compute an invariant representation not only for the whole image but especially for all parts of it which may contain objects and object parts and b) invariance to global transformations that are not affine (but are locally affine, that is, affine within the pooling range of some of the modules in the hierarchy)[3] Fig. 10 show examples of invariance and stability for wholes and parts. In the architecture of Fig. 1, each $\bigwedge$-module provides uniqueness, invariance and stability at different levels, over increasing ranges from bottom to top. Thus, in addition to the desired properties of invariance, stability and discriminability, these architectures match the hierarchical structure of the visual world and the need to retrieve items from memory at various levels of size and complexity. The results described

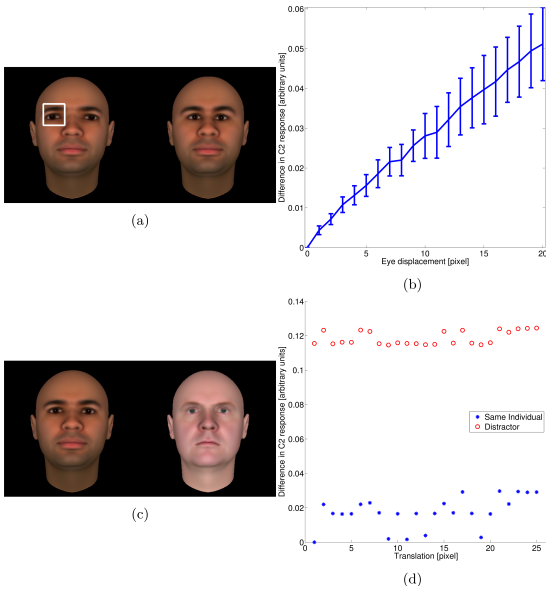

Fig. 3: *Empirical demonstration of the properties of invariance, stability and uniqueness of the hierarchical architecture (see Theorem 12) in a specific 2 layers implementation (HMAX). Inset (a) shows the reference image on the left and a deformation of it (the eyes are closer to each other) on the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

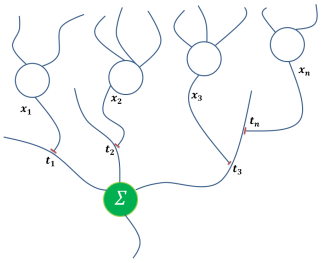

Fig. 4: *A neuron (green) can easily perform high-dimensional inner products between inputs on its dendritic tree and stored synapse weights.*

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

here are part of a general theory of hierarchical architectures which is beginning to take form (see [35, 48, 49, 50]) around the basic function of computing invariant representations.

The property of compositionality discussed above is related to the efficacy of hierarchical architectures vs. one-layer architectures in dealing with the problem of partial occlusion and the more difficult problem of clutter in object recognition. Hierarchical architectures are better at recognition in clutter than one-layer networks [51], because they provide signatures for image patches of several sizes and locations. However, hierarchical feedforward architectures cannot fully solve the problem of clutter. More complex (e.g. recurrent) architectures are likely needed for human-level recognition in clutter (see for instance [52, 53, 54]) and for other aspects of human vision. It is likely that much of the circuitry of visual cortex is required by these recurrent computations, not considered in this paper.

Visual Cortex

The theory described above effectively maps the computation of an invariant signature onto well-known capabilities of cortical neurons. A key difference between the basic elements of our digital computers and neurons is the number of connections: 3 vs. 10^3-10^4 synapses per cortical neuron. Taking into account basic properties of synapses, it follows that a single neuron can compute high-dimensional (10^3-10^4) inner products between input vectors and the stored vector of synaptic weights [55]. A natural scenario is then the following (see also Fig. 4). Consider an HW-module of "simple" and "complex" cells [31] looking at the image through a window defined by their receptive fields (see SI Appendix, section 1). Suppose that images of objects in the visual environment undergo affine transformations. During development—and more generally, during visual experience—a set of $|G|$ simple cells store in their synapses an image patch t^k and its transformations $g_1 t^k, ..., g_{|G|} t^k$—one per simple cell. This is done, possibly at separate times, for K different image patches t^k (templates), $k = 1, \cdots, K$. Each gt^k for $g \in G$ is a sequence of frames, literally a movie of image patch t^k transforming. There is a very *simple, general, and powerful way to learn* such unconstrained transformations. Unsupervised (Hebbian) learning is the main mechanism: for a "complex" cell to pool over several simple cells, the key is an unsupervised Foldiak-type rule: *cells that fire together are wired together.* At the level of complex cells this rule determines *classes of equivalence* among simple cells – reflecting observed *time correlations in the real world, that is transformations* of the image. Time continuity, induced by the Markovian physics of the world, allows associative labeling of stimuli based on their temporal contiguity.

Later, when an image is presented, the simple cells compute $\langle I, g_i t^k \rangle$ for $i = 1, ..., |G|$. The next step, as described above, is to estimate the one-dimensional probability distribution of such a projection, that is the distribution of the outputs of the simple cells. It is generally assumed that complex cells pool the outputs of simple cells. Thus a complex cell could compute $\mu_n^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + n\Delta)$ where σ is a smooth version of the step function ($\sigma(x) = 0$ for $x \leq 0$, $\sigma(x) = 1$ for $x > 0$) and $n = 1, ..., N$. Each of these N complex cells would estimate one bin of an approximated CDF (cumulative distribution function) for $P_{\langle I, t^k \rangle}$. Following the theoretical arguments above, the complex cells could compute, instead of an empirical CDF, one or more of its moments. $n = 1$ is the mean of the dot products, $n = 2$ corresponds to an energy model of complex cells [56]; very large n corresponds to a *max* operation. Conventional wisdom interprets available physiological data to suggest that simple/complex cells in V1 may be described in terms of energy models, but our alternative suggestion of empirical histogramming by sigmoidal nonlinearities with different offsets may fit the diversity of data even better.

As described above, a template and its transformed versions may be learned from unsupervised visual experience through Hebbian plasticity. Remarkably, our analysis and empirical studies[35] show that Hebbian plasticity, as formalized by Oja, can yield *Gabor-like tuning*—i.e., the templates that provide optimal invariance to translation and scale (see SI Appendix section 1)[4].

The localization condition (Equation 15) can also be satisfied by images and templates that are similar to each other. The result is invariance to class-specific transformations. This part of the theory is consistent with the existence of class-specific modules in primate cortex such as a face module and a body module [62, 63, 36]. It is intriguing that *the same localization condition* suggests *general Gabor-like templates for generic images* in the first layers of a hierarchical architectures and *specific, sharply tuned templates* for the last stages of the hierarchy[5]. This theory also fits physiology data concerning Gabor-like tuning in V1 and possibly in V4 (see [35]). It can also be shown that the theory, together with the hypothesis that storage of the templates takes place via Hebbian synapses, also predicts properties of the tuning of neurons in the face patch AL of macaque visual cortex [35, 64].

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

Discussion

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

[4] There is psychophysical and neurophysiological evidence that the brain employs such learning rules (e.g. [58, 60] and references therein). A second step of Hebbian learning may be responsible for wiring a complex cells to simple cells that are activated in close temporal contiguity and thus correspond to the same patch of image undergoing a transformation in time [57]. Simulations show that the system could be remarkably robust to violations of the learning rule's assumption that temporally adjacent images correspond to the same object [61]. The same simulations also suggest that the theory described here is qualitatively consistent with recent results on plasticity of single IT neurons and with experimentally-induced disruptions of their invariance [60].

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

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

across levels of analysis [35, 65]. At the computational level, it gives a unified account of *why* a range of seemingly different models have recently achieved impressive results on recognition tasks. HMAX [32, 66, 67], Convolutional Neural Networks [33, 34, 68, 69] and Deep Feedforward Neural Networks [44, 45, 46] are examples of this class of architectures—as is, possibly, the feedforward organization of the ventral steam. In particular, the theoretical framework of this paper may help explain the recent successes of hierarchical architectures of convolutional type on visual and speech recognition tests e.g. [45, 44]). At the algorithmic level, it motivates the development, now underway, of a new class of models for vision and speech which includes the previous models as special cases. At the level of biological implementation, its characterization of the optimal tuning of neurons in the ventral stream is consistent with the available data on Gabor-like tuning in V1 ([35]) and the more specific types of tuning in higher areas such as in faces patches.

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

ACKNOWLEDGMENTS. We would like to thank the McGovern Institute for Brain Research for their support. We would also like to thank for detailed and helpful comments on the manuscript Steve Smale, Stephane Mallat, Marco Cuturi, Robert Desimone, Jake Bouvrie, Charles Cadieu, Ryan Rifkin, Andrew Ng, Terry Sejnowski. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by the Eugene McDermott Foundation.

Supplementary Information

0. Setup and Definitions

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

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

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

Remarks:

1. The k-th component of the signature associated with a *simple-complex* module is (see Equation **[12]**) $\mu_n^k(I) = \frac{1}{|G_0|}\sum_{g\in G_0}\eta_n\Big(\left\langle gI, t^k\right\rangle\Big)$ where the functions η_n are such that $Ker(\eta_n) = \{0\}$: in words, the empirical histogram estimated for $\left\langle gI, t^k\right\rangle$ does not take into account the 0 value, since it does not carry any information about the image patch. The functions η_n are also assumed to be positive and invertible.
2. Images I are inputs to the modules of later one and have a maximum total possible support corresponding to a bounded region $B \subseteq \mathbb{R}^2$, which we refer to as the *visual field*, and which corresponds to the spatial pooling range of the module at the top of the hierarchy of Figure 1 in the main text. *Neuronal images* also written as I are inputs to the modules in higher layers and are usually supported in a higher dimensional space $\mathbb{R}^d$, corresponding to the signature components provided by lower layers modules; *isolated objects* are images with support contained in the pooling range of one of the modules at an intermediate level of the hierarchy. We use the notation $\nu(I), \mu(I)$ respectively for the simple responses $\left\langle gI, t^k\right\rangle$ and for the complex response $\mu_n^k(I) = \frac{1}{|G_0|}\sum_{g\in G_0}\eta_n\Big(\left\langle gI, t^k\right\rangle\Big)$. To simplify the notation we suppose that the center of the support of $\mu_\ell(I)$ coincides with the center of the pooling range.
3. The domain of the dot products $\left\langle gI, t^k\right\rangle$ corresponding to templates and to simple cells is in general different from the domain of the pooling $\sum_{g\in G_0}$. We will continue to use the commonly used term *receptive field* – even if it mixes these two domains.
4. The main part of the theory characterizes properties of the basic HW module – which computes the components of an invariant signature vector from an image patch within its receptive field.
5. It is important to emphasize that the *basic module is always the same* throughout the paper. We use different mathematical tools, including approximations, to study under which conditions (e.g. localization or linearization, see end of section 1) the signature computed by the module is invariant or approximatively invariant.
6. The pooling $\sum_{g\in G_0}$ is effectively over a *pooling window* in the group parameters. In the case of 1D scaling and 1D translations, the pooling window corresponds to an interval, e.g. $[a^j, a^{j+k}]$, of scales and an interval, e.g. $[-\bar{x}, \bar{x}]$, of x translations, respectively.
7. All the results in this paper are valid in the case of a discrete or a continuous compact group: in the first case we have a sum over the transformations, in the second an integral over the Haar measure of the group. In the following, for convenience, the theorems are proved in the continuous setting.
8. Normalized dot products also eliminate the need of the explicit computation of the determinant of the Jacobian for affine transformations (which is a constant and is simplified dividing by the norms) assuring that $\langle AI, At\rangle = \langle I, t\rangle$, where A is an affine transformation.

1. Basic Module

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

In this section we study the signature given by

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Theorem 2. *(Cramer-Wold, [40]) For any pair P, Q of probability distributions on $\mathbb{R}^d$, we have that $P = Q$ if and only if $\mathcal{E} = \mathbb{S}(\mathbb{R}^d)$.*

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

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

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

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I')) d\lambda(t), \quad \forall I, I' \in \mathcal{X}, \qquad \mathbf{[3]}$$

where d_μ is a metric on histograms induced by d_0.

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

$$\hat{d}_K(P_I, P_{I'}) = \frac{1}{K} \sum_{k=1}^{K} d_\mu(\mu^k(I), \mu^k(I')), \qquad \mathbf{[4]}$$

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

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\| \mu^k(I) - \mu^k(I) \right\|_{\mathbb{R}^N}. \qquad \mathbf{[5]}$$

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

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

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon, \qquad [6]$$

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

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

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

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

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

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

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

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

$$\mu_n^k(I) = \frac{1}{|G|}\sum_{g\in G} \eta_n\Big(\big\langle gI, t^k\big\rangle\Big), \qquad [7]$$

for G finite group, or equivalently

$$\mu_n^k(I) = \int_G dg\ \eta_n\Big(\big\langle gI, t^k\big\rangle\Big) = \int_G dg\ \eta_n\Big(\big\langle I, g^{-1}t^k\big\rangle\Big), \quad [8]$$

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

$$\mu_n^k(\bar{g}I) = \mu_n^k(I), \quad \forall \bar{g} \in G, I \in \mathcal{X}. \qquad [9]$$

Stability. With $\Sigma(I) \in \mathbb{R}^{NK}$ denoting as usual the signature of an image, and $d(\Sigma(I), \Sigma(I'))$, $I, I' \in \mathcal{X}$, a metric, we say that a *signature Σ is stable if it is Lipschitz continuous* (see [48]), that is

$$d(\Sigma(I), \Sigma(I')) \leq L\left\|I - I'\right\|_2, \quad L > 0, \ \forall I, I' \in \mathcal{X}. \qquad [10]$$

In our setting we let

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

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

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

$$d(\Sigma(I), \Sigma(I')) \leq \left\|I - I'\right\|_2, \qquad [11]$$

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

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

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

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

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

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

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

from which we obtain [**10**] summing over all K components and dividing by $1/K$. In particular if $NL_\eta \leq 1$ the map is non expansive and summing each component we have eq. [**11**]. Q.E.D.

The above result shows that the stability of the empirical signature

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

provided with the metric [**4**] (together with [**5**]) holds for nonlinearities with Lipschitz constants L_{η_n} such that $Nmax_n(L_{\eta_n}) < 1$.

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

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

Partially Observable Groups. This section outlines invariance, uniqueness and stability properties of the signature obtained in the case in which transformations of a group are observable only within a *window* "over" the orbit. The term POG (Partially Observable Groups) emphasizes the properties of the group – in particular associated invariants – as seen by an observer (e.g. a neuron) looking through a window at a part of the orbit. Let G be a finite group and $G_0 \subseteq G$ a subset (note: G_0 is not usually a subgroup). The subset of transformations G_0 can be seen as the set of transformations that can be *observed* by a window on the orbit that is the transformations that correspond to a part of the orbit. A *local* signature associated to the partial observation of G can be defined considering

$$\mu_n^k(I) = \frac{1}{|G_0|} \sum_{g \in G_0} \eta_n\Big(\Big\langle gI, t^k \Big\rangle \Big), \qquad [12]$$

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

$$\mu_n^k(I) = \frac{1}{V_0} \int_{G_0} \eta_n\Big(\Big\langle gI, t^k \Big\rangle \Big) dg, \quad V_0 = \int_{G_0} dg. \qquad [13]$$

Note that the constant V_0 normalizes the Haar measure, restricted to G_0, so that it defines a probability distribution. The latter is the distribution of the images subject to the group transformations which are observable, that is in G_0. The above definitions can be compared to definitions [**7**] and [**8**] in the fully observable groups case. In the next sections we discuss the properties of the above signature. While stability and uniqueness follow essentially from the analysis of the previous section, invariance requires developing a new analysis.

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

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

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

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

$$\mu_n^k(I) = \frac{1}{V_0} \int_{G_0} dg\, \eta_n\Big(\Big\langle gI, t^k \Big\rangle \Big). \qquad [14]$$

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

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

$$\mu_n^k(I) = \mu_n^k(\bar{g}I) \;\Leftrightarrow\; \Big\langle gI, t^k \Big\rangle = 0,\ \forall g \in G/(G_0 \cap \bar{g}G_0),$$
$$\Big\langle gI^k, t \Big\rangle \neq 0,\ \forall g \in G_0 \cap \bar{g}G_0. \qquad [15]$$

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

$$\begin{aligned} 0 &= \int_{G_0} dg\, \eta_n\Big(\big(\Big\langle gI, t^k \Big\rangle \big) - \eta_n\big(\Big\langle g\bar{g}I, t^k \Big\rangle \big) \Big) \\ &= \int_{G_0 \Delta \bar{g}G_0} dg\, \eta_n\big(\Big\langle gI, t^k \Big\rangle \big) \qquad [16] \\ &= \int_{G/(G_0 \cap \bar{g}G_0)} dg\, \eta_n\big(\Big\langle gI, t^k \Big\rangle \big) \end{aligned}$$

where Δ is the symbol for symmetric difference ($A \Delta B = (A \cup B)/(A \cap B)$ A, B *sets*) and the last equality holds if

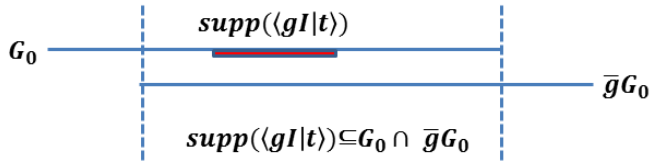

Fig. 5: *Necessary and sufficient condition for local invariance: if the support of $\langle gI, t \rangle$ is sufficiently localized it will be completely contained in the pooling interval even if the image is group shifted, or, equivalently (as shown in the Figure), if the pooling interval is group shifted by the same amount.*

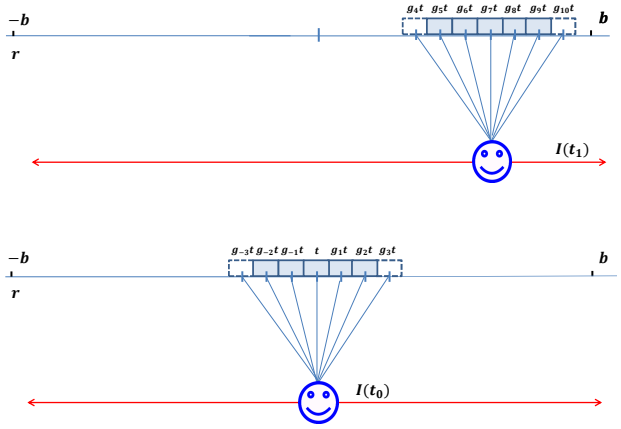

Fig. 6: *An HW-module pooling the dot products of transformed templates with the image. The input image I is shown centered on the template t; the same module is shown above for a group shift of the input image, which now localizes around the transformed template $g_7 t$. Images and templates satisfy the localization condition $\langle I, T_x t \rangle \neq 0$, $|x| > a$ with $a = 3$. The interval $[-b, b]$ indicates the pooling window. The shift in x shown in the Figure is a special case: the reader should consider the case in which the transformation parameter, instead of x, is for instance rotation in depth.*

$supp(\langle gI, t^k \rangle) \subseteq G_0$. Since the functions η_n are positive and bijective, eq. **[16]** implies $\langle gI, t^k \rangle = 0,\ g \in G/(G_0 \cap \bar{g}G_0)$. The inverse implication is proved by simply inverting the chain of equalities. See Figure 5 for a visual explanation. Q.E.D.
Condition in eq. **[16]** is a *localization* condition on the product of the transformed image and the template (see Figure 17 for a pictorial intuitive example in the case of translation group). In the next paragraph we will see how localization conditions for scale and translation transformations implies a specific form of the templates.

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

In the following it is helpful to assume that all images I and templates t are strictly contained in the range of translation or scale pooling, P, since image components outside it are not measured. We will consider images I *restricted to* P: for translation this means that the support of I is contained in P, for scaling, since $g_s I = I(sx)$ and $\widehat{I(sx)} = (1/s)\hat{I}(\omega/s)$ (where $\hat{\cdot}$ indicates the Fourier transform), assuming a scale pooling range of $[s_m, s_M]$, implies a range $[\omega_m^I, \omega_M^I],\ [\omega_m^t, \omega_M^t]$ (m and M indicates maximum and minimum) of spatial frequencies for the maximum support of I and t. As we will see because of Theorem 5 *invariance to translation requires spatial localization of images and templates* and less obviously *invariance to scale requires bandpass properties of images and templates*. Thus images and templates are assumed to be localized from the outset in either space or frequency. The corollaries below show that a stricter localization condition is needed for invariance and that this condition determines the form of the template. Notice that in our framework images and templates are bandpass because of being zero-mean. Notice that, in addition, neural "images" which are input to the hierarchical architecture are spatially bandpass because of retinal processing.
We now state the result of Theorem 5 for one dimensional signals under the translation group and – separately – under the dilation group.
Let $I, t \in L^2(\mathbb{R})$, $(\mathbb{R}, +)$ the one dimensional locally compact group of translations and $T_x : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ a unitary representation of the translation operator. Let, e.g., $G_0 = [-b, b],\ b > 0$ and suppose $supp(t) \subseteq supp(I) \subseteq [-b, b]$. Further suppose $supp(\langle T_x I, t \rangle) \subseteq [-b, b]$. Then eq. **[15]** specializes to

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

$$\mu_n^k(I) = \mu_n^k(T_x I),\ \forall x \in [0, \bar{x}] \ \Leftrightarrow\ \langle T_x I, t \rangle \neq 0,\ \forall x \in [-b+\bar{x}, b]. \quad \textbf{[17]}$$

with $\bar{x} > 0$.

Similarly let $G = (\mathbb{R}^+, \cdot)$ be the one dimensional locally compact group of dilations and denote with $D_s : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ a unitary representation of the dilation operator. Let $G_0 = [1/S, S],\ S > 1$ and suppose $supp(\langle D_s I, t \rangle) \subseteq [1/S, S]$. Then eq. **[15]** gives

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

$$\mu_n^k(I) = \mu_n^k(D_s I), s \in [1, \bar{s}] \ \Leftrightarrow\ \langle D_s I, t \rangle \neq 0,\ \forall s \in [\frac{\bar{s}}{S}, S]. \quad \textbf{[18]}$$

with $S > 1$.
Localization conditions of the support of the dot product for translation and scale are depicted in Figure 7,a),b).
As shown by the following Lemma 1 Eq. **[17]** and **[18]** gives interesting conditions on the supports of t and its Fourier transform $\hat{t}$. For translation, the corollary is equivalent to zero overlap of the compact supports of I and t. In particular using Theorem 5, for $I = t$, the maximal invariance implies the following localization conditions on t

$$\langle gt, t \rangle = 0 \quad g \notin G_L \subseteq G \quad \textbf{[19]}$$

which we call self-localization. For $1D$ translations it has the simple form $\langle T_x t, t \rangle = 0\ |x| > a,\ a > 0$.
For scaling we consider the support of the Fourier transforms of I and t. The Parseval theorem allows to rewrite the dot product $\langle D_s I, t \rangle$ which is in $L^2(\mathbb{R}^2)$ as $\left\langle \widehat{D_s I}, \hat{t} \right\rangle$ in the Fourier domain.
In the following we suppose that the support of $\hat{t}$ and $\hat{I}$ is respectively $[\omega_m^t, \omega_M^t]$ and $[\omega_m^I, \omega_M^I]$ where $\omega_m^{t,I}$ could be very close to zero (images and templates are supposed to be zero-mean) but usually are bigger then zero.
Note that the effect of scaling I with (typically $s = 2^j$ with $j \leq 0$) is to change the support as $supp(\widehat{D_s I}) = s(supp(\hat{I}))$. This change of the support of $\hat{I}$ in the dot product $\left\langle \widehat{D_s I}, \hat{t} \right\rangle$ gives non trivial conditions on the intersection with the support of $\hat{t}$ and therefore on the localization w.r.t. the scale invariance. We have the following Lemma:

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

$$supp(t) \subseteq [-b - \bar{x}, b] - supp(I),\ I \in \mathcal{X}. \quad \textbf{[20]}$$

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

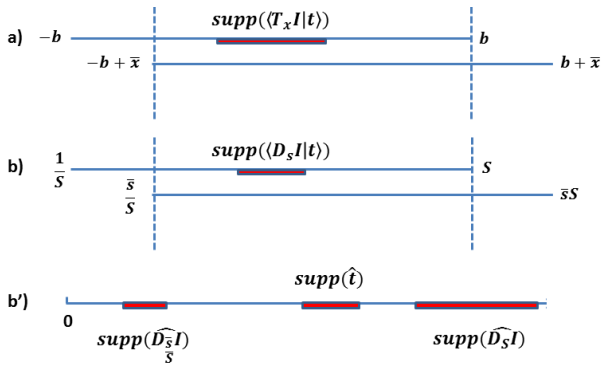

Fig. 7: *a), b): if the support of the dot product between the image and the template is contained in the intersection between the pooling range and the group translated (a) or dilated (b) pooling range the signature is invariant. In frequency condition b) becomes b'): when the Fourier supports of the dilated image and the template do not intersect their dot product is zero.*

quency ω

$$supp(\hat{t}) \subseteq [-\omega_t - \Delta_t^*, -\omega_t + \Delta_t^*] \cup [\omega_t - \Delta_t^*, \omega_t + \Delta_t^*]$$
$$\Delta_t^* = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S}, \;\; \omega_t = \frac{\omega_M^t - \omega_m^t}{2}. \qquad [21]$$

Proof:
To prove that $supp(t) \subseteq [-b + \bar{x}, b] - supp(I)$ note that eq. **[17]** implies that $supp(\langle T_x I, t\rangle) \subseteq [-b + \bar{x}, b]$ (see Figure 7, a)). Being $supp(\langle T_x I, t\rangle) = supp(I * t) \subseteq supp(I) + supp(t)$ we have $supp(t) \subseteq [-b - \bar{x}, b] - supp(I)$.
To prove the condition in eq. **[21]** note that eq. **[18]** is equivalent in the Fourier domain to

$$\langle D_s I, t\rangle = \Big\langle \widehat{D_s I}, \hat{t} \Big\rangle = \frac{1}{s} \int d\omega \; \hat{I}\big(\frac{\omega}{s}\big)\hat{t}(\omega) \neq 0 \;\; \forall s \in [\frac{\bar{s}}{S}, S] \qquad [22]$$

The situation is depicted in Fig. 7 b') for S big enough: in this case in fact we can suppose the support of $\widehat{D_{\bar{s}/S} I}$ to be on an interval on the left of that of $supp(\hat{t})$ and $\widehat{D_S I}$ on the right; the condition $supp(\Big\langle \widehat{D_s I}, \hat{t} \Big\rangle) \subseteq [\bar{s}/S, S]$ is in this case equivalent to

$$\omega_M^I \frac{\bar{s}}{S} < \omega_m^t, \;\; \omega_M^t < \omega_m^I S \qquad [23]$$

which gives

$$\Delta_t^* = Max(\Delta_t) \equiv Max\Big(\frac{\omega_M^t - \omega_m^t}{2}\Big) = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S} \qquad [24]$$

and therefore eq. **[21]**.Q.E.D.
Note that for some $s \in [\bar{s}/S, S]$ the condition that the Fourier supports are disjoint is only sufficient and not necessary for the dot product to be zero since cancelations can occur. However to have $\Big\langle \widehat{D_s I}, \hat{t} \Big\rangle = 0$ on a continuous interval of scales, (unless some pathological examples of the function I) implies disjointness of the supports, since $\hat{I}(\omega/s) \neq \hat{I}(\omega/s')$, $s \neq s'$ unless I has constant spectrum in the interval $[\bar{s}/S, S]$. A similar reasoning is valid for the translation case.

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

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

Proof:
We illustrate the statement of the theorem with a simple example. In the case of translations suppose, e.g., $supp(I) = [-b, b']$, $supp(t) = [-a, a]$, $a \leq b' \leq b$. Eq. **[20]** reads

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

which gives the condition $-a \geq -b + b' + \bar{x}$, i.e. $\bar{x}^{max} = b - b' - a$; thus, for any fixed b, b' the smaller the template support $2a$ in space, the greater is translation invariance.
Similarly, in the case of dilations, increasing the range of invariance $[1, \bar{s}]$, $\bar{s} > 1$ implies a decrease in the support of $\hat{t}$ as shown by eq. **[24]**; in fact noting that $|supp(\hat{t})| = 2\Delta_t$ we have

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

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

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

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

$$\Delta x \cdot \Delta\omega \geq 1. \qquad [25]$$

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

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

$$\psi_{x_0,\omega_0}(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}, \; \sigma_x \in \mathbb{R}^+, \; \omega_0 \in \mathbb{R} \qquad [26]$$

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

$$\langle T_x I, t\rangle = e^{-\frac{x^2}{\sigma_x^2}}, \quad \langle D_s I, t\rangle = e^{-\frac{s^2}{\sigma_s^2}}, \; \sigma_x, \sigma_s \in \mathbb{R} \qquad [27]$$

we have

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

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

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

$$t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}. \qquad [28]$$

Remarks

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

$$\int \psi(x)dx = 0 \qquad [29]$$

and

$$||\psi(x)|| = 1. \qquad [30]$$

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

$$\psi_{u,s}(x) = \frac{1}{s^{\frac{1}{2}}}\psi(\frac{x-u}{s}). \qquad [31]$$

Under certain conditions (in particular, the Heisenberg boxes associated with each wavelet must together cover the space-frequency plane) the Gabor wavelet family becomes a Gabor wavelet frame.
2. Optimal self-localization of the templates (which follows from localization), when valid simultaneously for space and scale, is also equivalent to Gabor wavelets. If they are a frame, full information can be preserved in an optimal quasi invariant way.

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

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

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

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

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

$$\begin{aligned} T_r(I) &= T_0(I) + \Big(\frac{dT}{dr}\Big)_{r=0}(I)r + R(I) \qquad [32] \\ &= I + \Big(\frac{dT}{dr}\Big)_{r=0}(I)r + R(I) \\ &= I + J^I(I)r + R(I) = [e + rJ^I](I) + R(I) \\ &= L_r^I(I) + R(I) \end{aligned}$$

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

$$\langle T_r I, t\rangle \approx \left\langle L_r^I I, t\right\rangle, \qquad [33]$$

If the above linearization holds, we have the following:

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

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

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

$$\begin{aligned} \mu_n^k(T_{\bar{r}}I) &= \int_P dr \ \eta_n(\langle T_r T_{\bar{r}} I, t\rangle) = \int_P dr \ \eta_n(\left\langle L_r^I L_{\bar{r}}^I I, t\right\rangle) \\ &= \int_P dr \ \eta_n(\left\langle L_{r+\bar{r}}^I I, t\right\rangle) = \mu_n^k(I) \end{aligned}$$

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

As an example, consider the transformation induced on the image plane by rotation in depth of a face: it can be decomposed into piecewise linear approximations around a small number of key templates, each one corresponding to a specific $3D$ rotation of a template face. Each key template corresponds to a complex cell containing as (simple cells) a number of observed transformations of the key template within a small range of rotations. Each key template corresponds to a different signature which is invariant only for rotations around its center. Notice that the form of the linear approximation or the number of key templates needed does not affect the algorithm or its implementation. The templates learned are used

in the standard dot-product-and-pooling module. The choice of the key templates – each one corresponding to a complex cell, and thus to a signature component – is not critical, as long as there are enough of them. For one parameter groups, the key templates correspond to the knots of a piecewise linear spline approximation. Optimal placement of the centers – if desired – is a separate problem that we leave aside for now.

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

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

Locally Compact Groups and Partially Observable Compact Groups: consider now the POG situation in which the pooling is over a subset of the group: (the POG case always applies to Locally Compact groups (LCG) such as translations). As shown before, a complex cell is partially invariant **if** the value of the dot-product between a template and its shifted template under the group falls to zero fast enough with the size of the shift relative to the extent of pooling.

In the POG and LCG case, such partial invariance holds over a restricted range of transformations if the templates and the inputs have a *localization* property that implies wavelets for transformations that include translation and scaling.

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

Remarks:

1. We are interested in two main cases of POG invariance:
   - partial invariance *simultaneously* to translations in x, y, scaling and possibly rotation in the image plane. This should apply to "generic" images. The signatures should ideally preserve full, locally invariant information. This first regime is ideal for the first layers of the multilayer network and may be related to Mallat's scattering transform, [48]. We call the sufficient condition for for LCG invariance here, *localization*, and in particular, *self-localization* given by Equation **[19]**.
   - partial invariance to linear transformations for a subset of all images. This second regime applies to high-level modules in the multilayer network specialized for specific classes of objects and non-group transformations. The condition that is sufficient here for LCG invariance is given by Theorem 5 which applies only to a specific class of I. We prefer to call it *sparsity* of the images with respect to a set of templates.
2. For classes of images that are sparse with respect to a set of templates, the localization condition does not imply wavelets. Instead it implies templates that are
   - similar to a class of images so that $\langle I, g_0 t^k \rangle \approx 1$ and
   - complex enough to be "noise-like" in the sense that $\langle I, g t^k \rangle \approx 0$ for $g \neq g_0$.
3. Templates must transform similarly to the input for approximate invariance to hold. This corresponds to the assumption of a class-specific module and of a *nice object class* [11, 36].
4. For the localization property to hold, the image must be similar to the key template or contain it as a diagnostic feature (a sparsity property). It must be also quasi-orthogonal (highly localized) under the action of the local group.
5. For a general, non-group, transformation it may be impossible to obtain invariance over the full range with a single signature; in general several are needed.
6. It would be desirable to derive a formal characterization of the error in local invariance by using the standard module of dot-product-and-pooling, equivalent to a complex cell. The above arguments provide the outline of a proof based on local linear approximation of the transformation and on the fact that a local linear transformation is a LCG.

2. Hierarchical Architectures

So far we have studied the invariance, uniqueness and stability properties of signatures, both in the case when a whole group of transformations is observable (see **[7]** and **[8]**), and in the case in which it is only partially observable (see **[12]** and **[13]**). We now discuss how the above ideas can be iterated to define a multilayer architecture. Consider first the case when G is finite. Given a subset $G_0 \subset G$, we can associate a *window* gG_0 to each $g \in G$. Then, we can use definition **[12]** to define for each window a signature $\Sigma(I)(g)$ given by the measurements,

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

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

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

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

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

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

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

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

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

$$\mu_\ell^{n,k}(I)(g) = \frac{1}{|G_\ell|}\sum_{\bar{g}\in gG_\ell}\eta_n\left(\nu_\ell^k(I)(\bar{g})\right), \qquad [34]$$

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

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

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

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

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

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

$$\mu_\ell^{n,k}(I)(g) = \frac{1}{V_\ell}\int d\bar{g}K_\ell(\bar{g}^{-1}g)\eta_n\left(\nu_\ell^k(I)(\bar{g})\right),\ g \in G \qquad [36]$$

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

$$\nu_\ell^k(I)(g) = \left\langle \mu_{\ell-1}(I), gt_\ell^k\right\rangle, \quad g \in G \qquad [37]$$

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

Remark Note that eq. [36] can be written as:

$$\mu_\ell^{n,k}(I) = K_\ell * \eta_n(\nu_\ell^k(I)) \qquad [38]$$

where $*$ is the group convolution.

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

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

In the following we show the covariance property for the $\mu_1^{n,k}$ response (see Fig. 8). An inductive reasoning then can be applied for higher order responses. We assume that the architecture is isotropic in the relevant covariance dimension (this implies that all the modules in each layer should be identical with identical templates) and that there is a continuum of modules in each layer.

Proposition 9. *Let G a locally compact group and $\bar{g} \in G$. Let $\mu_1^{n,k}$ as defined in 36. Then $\mu_1^{n,k}(\tilde{g}I)(g) = \mu_1^{n,k}(I)(\tilde{g}^{-1}g)$.*

Proof:
Using the definition 36 we have

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

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

Remarks

1. The covariance property described in proposition 9 can be stated equivalently as $\mu_1^{n,k}(I)(g) = \mu_1^{n,k}(\bar{g}I)(\bar{g}g)$. This last expression has a more intuitive meaning as shown in Fig. 8.
2. The covariance property described in proposition 9 holds both for abelian and non-abelian groups. However the group average on templates transformations in definition of eq. 36 is crucial. In fact, if we define the signature averaging on the images we do not have a covariant response:

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

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

Fig. 8: *Covariance: the response for an image I at position g is equal to the response of the group shifted image at the shifted position.*

3. With respect to the range of invariance, the following property holds for multilayer architectures in which the output of a layer is defined as covariant if it transforms in the same way as the input: for a given transformation of an image or part of it, the signature from complex cells at a certain layer is either invariant or covariant with respect to the group of transformations; if it is covariant there will be a higher layer in the network at which it is invariant (more formal details are given in theorem 11), assuming that the image is contained in the visual field. This property predicts a *stratification* of ranges of invariance in the ventral stream: invariances should appear in a sequential order meaning that smaller transformations will be invariant before larger ones, in earlier layers of the hierarchy (see [13]).

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

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

$$\mu_\ell(I) = \mu_\ell(\bar{g}I) \;\Leftrightarrow\; \begin{array}{l} \langle g\mu_{\ell-1}(I), t\rangle = 0,\ g \in G/(G_\ell \cap \bar{g}G_\ell), \\ \langle g\mu_{\ell-1}(I), t\rangle \neq 0,\ g \in G_\ell \cap \bar{g}G_\ell. \end{array} \quad [39]$$

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

We can now state the following:

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

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

The proof follows from the observation that the pooling range over the group is a bigger and bigger subset of G with growing layer number, in other words, there exists a layer such that the image and its transformations are within the pooling range at that layer (see Fig. 9). This is clear since for any $\bar{g} \in G$ the nested sequence

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

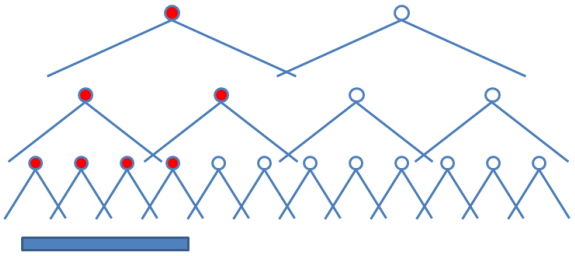

Fig. 9: *An image I with a finite support may or may not be fully included in the receptive field of a single complex cell at layer n (more in general the transformed image may not be included in the pooling range of the complex cell). However there will be a higher layer such that the support of its neural response is included in the pooling range of a single complex cell.*

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

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

being $supp(\left\langle g\mu_{\bar{\ell}-1}(I), t\right\rangle) \subseteq G$. Details are reported in [35].

Property 3: stability. Using the definition of stability given in [**11**], we can formulate the following theorem characterizing stability for the complex response:

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

$$\left\|\mu_\ell(I) - \mu_\ell(I')\right\| \leq \left\|I - I'\right\|,\ \forall\ \ell,\ \forall\ I, I' \in \mathcal{X}. \quad [40]$$

The proof follows from a repeated application of the reasoning done in Theorem 11. See details in [35].

Comparison with stability defined by Mallat [48]. The same definition of stability we use (Lipschitz continuity) was recently given by [48], in a related context. Let $I, I' \in L^2(\mathbb{R}^2)$ and $\Phi : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ a representation. Φ is stable if it is Lipschitz continuous with Lipschitz constant $L \leq 1$, i.e., is a non expansive map:

$$\left\|\Phi(I) - \Phi(I')\right\|_2 \leq \left\|I - I'\right\|_2,\ \ \forall\ I, I' \in L^2(\mathbb{R}^2). \quad [41]$$

In particular in [48] the author is interested in stability of group invariant scattering representations to the action of small diffeomorphisms close to translations. Consider transformations of the form $I'(\mathbf{x}) = \mathbf{L}_\tau \mathbf{I}(\mathbf{x}) = \mathbf{I}(\mathbf{x} - \tau(\mathbf{x}))$ (which can be though as small diffeomorphic transformations close to translations implemented by a displacement field $\tau : \mathbb{R}^2 \to \mathbb{R}^2$). A translation invariant operator Φ is said to be Lipschitz continuous to the action of a $C^2(\mathbb{R}^2)$ diffeomorphisms if for any compact $\Omega \subseteq \mathbb{R}^2$ there exists C such that for all $I \in L^2(\mathbb{R}^2)$ supported in $\Omega \subseteq \mathbb{R}^2$ and $\tau \in C^2(\mathbb{R}^2)$

$$\begin{array}{l} \left\|\Phi(I) - \Phi(L_\tau I)\right\|_2 \leq \\ \leq C \left\|I\right\|_2 \left(sup_{\mathbf{x}\in\mathbb{R}^2}|\nabla\tau(\mathbf{x})| + sup_{\mathbf{x}\in\mathbb{R}^2}|H\tau(\mathbf{x})|\right) \end{array} \quad [42]$$

where H is the Hessian and C a positive constant.
Condition [**42**] is a different condition then that in eq. [**40**] since it gives a Lipschitz bound for a diffeomorphic transformation at each layer of the scattering representation.
Our approach differs in the assumption that small (close to identity) diffeomorphic transformations can be well approximated, at the first layer, as locally affine transformations or, in the limit, as local translations which therefore falls in the

POG case. This assumption is substantiated by the following reasoning in which any smooth transformation is seen as parametrized by the parameter t (the r parameter of the T_r transformation in section 1), which can be thought as time. Let $T \subseteq \mathbb{R}$ be a bounded interval and $\Omega \subseteq \mathbb{R}^N$ an open set and let $\Phi = (\Phi_1, ..., \Phi_N) : T \times \Omega \to \mathbb{R}^N$ be $\mathcal{C}_2$ (twice differentiable), where $\Phi(0, .)$ is the identity map. Here $\mathbb{R}^N$ is assumed to model the image plane, intuitively we should take $N = 2$, but general values of N allow our result to apply in subsequent, more complex processing stages, for example continuous wavelet expansions, where the image is also parameterized in scale and orientation, in which case we should take $N = 4$. We write (t, x) for points in $T \times \Omega$, and interpret $\Phi(t, x)$ as the position in the image at time t of an observed surface feature which is mapped to $x = \Phi(0, x)$ at time zero. The map Φ results from the (not necessarily rigid) motions of the observed object, the motions of the observer and the properties of the imaging apparatus. The implicit assumption here is that no surface features which are visible in Ω at time zero are lost within the time interval T. The assumption that Φ is twice differentiable reflects assumed smoothness properties of the surface manifold, the fact that object and observer are assumed massive, and corresponding smoothness properties of the imaging apparatus, including eventual processing. Now consider a closed ball $B \subset \Omega$ of radius $\delta > 0$ which models the aperture of observation. We may assume B to be centered at zero, and we may equally take the time of observation to be $t_0 = 0 \in T$. Let

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

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

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

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

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

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

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

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

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

whence

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

Q.E.D.

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

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

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

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

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

Theorem 13 and corollary 14 gives a precise mathematical motivation for the assumption that any sufficiently smooth (at least twice differentiable) transformation can be approximated in an enough small compact set with a group transformation (e.g. translation), thus allowing, based on eq. 11, stability w.r.t. small diffeomorphic transformations.

Approximate Factorization: hierarchy. In the first version of [35] we conjectured that a signature invariant to a group of transformations could be obtained by factorizing in successive layers the computation of signatures invariant to a subgroup of the transformations (e.g. the subgroup of translations of the affine group) and then adding the invariance w.r.t. another subgroup (e.g. rotations). While factorization of invariance ranges is possible in a hierarchical architecture (theorem 11), it can be shown that in general the factorization in successive layers for instance of invariance to translation followed by invariance to rotation (by subgroups) is impossible (see [35]). However, approximate factorization is possible under the same conditions of the previous section. In fact, a transformation that can be linearized piecewise can always be performed in higher layers, on top of other transformations, since the global group structure is not required but weaker smoothness properties are sufficient.

Why Hierarchical architectures: a summary.

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

2. *Compositionality.* A hierarchical architecture provides signatures of larger and larger patches of the image in terms of lower level signatures. Because of this, it can access memory in a way that matches naturally with the linguistic ability to describe a scene as a whole and as a hierarchy of parts.
3. *Approximate factorization.* In architectures such as the network sketched in Fig. 1 in the main text, approximate invariance to transformations specific for an object class can be learned and computed in different stages. This property may provide an advantage in terms of the sample complexity of multistage learning [15]. For instance, approximate class-specific invariance to pose (e.g. for faces) can be computed on top of a translation-and-scale-invariant representation [36]. Thus the implementation of invariance can, in some cases, be "factorized" into different steps corresponding to different transformations. (see also [16, 17] for related ideas).

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

3. Synopsis of Mathematical Results

List of Theorems

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

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

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

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

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

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

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

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

  implies (for compact groups without any additional property) that the signature components $\mu_n^k(I) = \frac{1}{|G|} \sum_{g \in G} \eta_n \Big(\left\langle I, gt^k \right\rangle \Big)$, calculated on templates transformations are invariant that is $\mu_n^k(I) = \mu_n^k(\bar{g} I)$.
- Invariance for Partially Observable Groups (observed through a window) is equivalent to condition in eq. **[19]** on the dot product between image and template)

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

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

- Condition in **[19]** is equivalent to a localization or sparsity property of the dot product between image and template ($\langle I, gt \rangle = 0$ for $g \notin G_L$). In particular

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

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

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

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

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

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

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

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

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

- Factorization.

  Proposition 8. *Invariance to separate subgroups of affine group cannot be obtained in a sequence of layers while factorization of the ranges of invariance can (because of covariance). Invariance to a smooth (non group) transformation can always be performed in higher layers, on top of other transformations, since the global group structure is not required.*

- Uniqueness of signature. **Conjecture:** *The neural image at the first layer is uniquely represented by the final signature at the top of the hierarchy and the means and norms at each layer.*

4. General Remarks on the Theory

1. The second regime of localization (sparsity) can be considered as a way to deal with situations that do not fall under the general rules (group transformations) by creating a series of exceptions, one for each object class.
2. Whereas the first regime "predicts" Gabor tuning of neurons in the first layers of sensory systems, the second regime predicts cells that are tuned to much more complex features, perhaps similar to neurons in inferotemporal cortex.
3. The *sparsity condition under the group* is related to properties used in associative memories for instance of the holographic type (see [8]). If the sparsity condition holds only for $I = t^k$ and for very small a then it implies strictly memory-based recognition.
4. The theory is memory-based. It also view-based. Even assuming 3D images (for instance by using stereo information) the various stages will be based on the use of 3D views and on stored sequences of 3D views.
5. The mathematics of the class-specific modules at the top of the hierarchy – with the underlying localization condition – justifies old models of viewpoint-invariant recognition (see [18]).
6. The remark on factorization of general transformations implies that layers dealing with general transformations can be on top of each other. It is possible – as empirical results by Leibo and Li indicate – that a second layer can improve the invariance to a specific transformation of a lower layer.
7. The theory developed here for vision also applies to other sensory modalities, in particular speech.
8. The theory represents a general framework for using representations that are invariant to transformations that are learned in an unsupervised way in order to reduce the sample complexity of the supervised learning step.
9. Simple cells (e.g. templates) under the action of the affine group span a set of positions and scales and orientations. The size of their receptive fields therefore spans a range. The pooling window can be arbitrarily large – and this does not affect selectivity when the CDF is used for pooling. A large pooling window implies that the signature is given to large patches and the signature is invariant to uniform affine transformations of the patches within the window. A hierarchy of pooling windows provides signature to patches and subpatches and more invariance (to more complex transformations).
10. Connections with the *Scattering Transform*.
    - Our theorems about optimal invariance to scale and translation implying Gabor functions (first regime) may provide a justification for the use of Gabor wavelets by Mallat [48], that does not depend on the specific use of the modulus as a pooling mechanism.
    - Our theory justifies several different kinds of pooling of which Mallat's seems to be a special case.
    - With the choice of the modulo as a pooling mechanisms, Mallat proves a nice property of Lipschitz continuity on diffeomorphisms. Such a property is not valid *in general* for our scheme where it is replaced by a hierarchical *parts and wholes* property which can be regarded as an approximation, as refined as desired, of the continuity w.r.t. diffeomorphisms.
    - Our second regime does not have an obvious corresponding notion in the scattering transform theory.
11. The theory characterizes under which conditions the signature provided by a HW module at some level of the hierarchy is invariant and therefore could be used for retrieving information (such as the label of the image patch) from memory. The simplest scenario is that signatures from modules at all levels of the hierarchy (possibly not the lowest ones) will be checked against the memory. Since there are of course many cases in which the signature will not be invariant (for instance when the relevant image patch is larger than the receptive field of the module) this scenario implies that the step of memory retrieval/classification is selective enough to discard efficiently the "wrong" signatures that do not have a match in memory. This is a nontrivial constraint. It probably implies that signatures at the top level should be matched first (since they are the most likely to be invariant and they are fewer) and lower level signatures will be matched next possibly constrained by the results of the top-level matches – in a way similar to *reverse hierarchies* ideas. It also has interesting implications for appropriate encoding of signatures to make them optimally quasi-orthogonal e.g. incoherent, in order to minimize memory interference. These properties of the representation depend on memory constraints and will be object of a future paper on memory modules for recognition.

5. Empirical support for the theory

Several computational vision models in recent literature can be considered instances of the theory described here. HMAX, trained convolutional networks, and the feedforward networks of N. Pinto et al. all consist of hierarchically stacked modules of simple and complex cells. However, only the most recent of these – variants of HMAX that incorporate invariances to complex transformations learned from video – have been designed with this theory explicitly in mind.

In [36], we showed that our approach of pooling over stored views of template faces undergoing the transformation can be used to recognize novel faces robustly to rotations in depth from a single example view. More recently, we applied the same idea to unconstrained face recognition benchmarks: Labeled Faces in the Wild and PubFig83, and showed that they yield a system that performs comparably to the state of the art with considerably less engineering.

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

The best-performing version of HMAX for generic object categorization is an improved version of [66] which scores 74% on the Caltech 101 dataset, competitive with the state-of-the-art for a single feature type. The original version achieved a near-perfect score on the UIUC car dataset. Another HMAX variant added a time dimension for action recognition [20], outperforming both human annotators and a state-of-the-art commercial system on a mouse behavioral phenotyping task. An HMAX model [67] was also shown to account for human performance in rapid scene categorization.

One of the observations that inspired our theory is that in convolutional architectures, random features perform nearly as well as features learned from objects [22, 23]. This includes models other than HMAX: [42] found that a convolutional

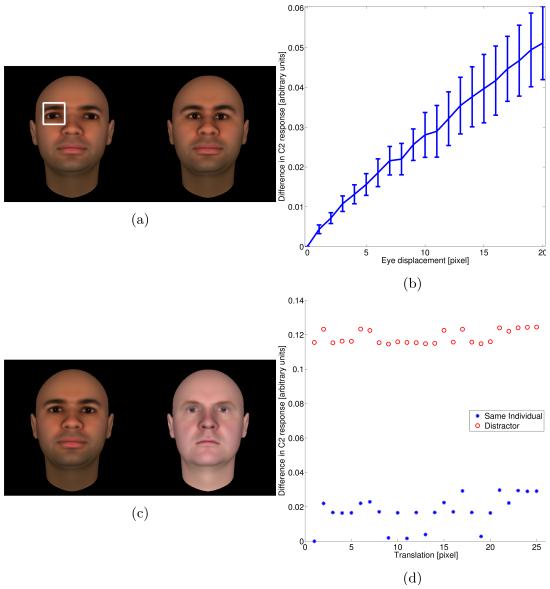

Fig. 10: *Empirical demonstration of the properties of invariance, stability and uniqueness of the hierarchical architecture (see Theorem 12) in a specific 2 layers implementation (HMAX). Inset (a) shows the reference image on the left and a deformation of it (the eyes are closer to each other) on the right; (b) shows that an HW-module in layer 1 whose receptive fields covers the left eye provides a signature vector (c_1) which is invariant to the deformation; in (c) an HW-module at layer 2 (c_2) whose receptive fields contain the whole face provides a signature vector which is (Lipschitz) stable with respect to the deformation. In all cases, the Figure shows just the Euclidean norm of the signature vector. Notice that the c_1 and c_2 vectors are not only invariant but also selective. Error bars represent ± 1 standard deviation. Two different images (d) are presented at various location in the visual field. The Euclidean distance between the signatures of a set of HW-modules at layer 2 with the same receptive field (the whole image) and a reference vector is shown in (e). The signature vector is invariant to global translation and discriminative (between the two faces). In this example the HW-module represents the top of a hierarchical, convolutional architecture. The images we used were 200×200 pixels*

network with randomized weights performed only 3% worse than the same network after training via backpropagation. [43] also found feature learning to be the least significant of several variables contributing to the performance of a hierarchical architecture.

A simple illustrative empirical demonstration of the HMAX properties of invariance, stability and uniqueness is in Fig. 10.

6.Unsupervised learning of the template orbit

While the templates need not be related to the test images (in the affine case), during development, the model still needs to observe the orbit of some templates. We conjectured that this could be done by unsupervised learning based on the temporal adjacency assumption [57, 27]. One might ask, do "errors of temporal association" happen all the time over the course of normal vision? Lights turn on and off, objects are occluded, you blink your eyes – all of these should cause errors. If temporal association is really the method by which all the images of the template orbits are associated with one another, why doesn't the fact that its assumptions are so often violated lead to huge errors in invariance?

The full orbit is needed, at least in theory. In practice we have found that significant scrambling is possible as long as the errors are not correlated. That is, normally an HW-module would pool all the $\langle I, g_i t^k \rangle$. We tested the effect of, for some i, replacing t^k with a different template $t^{k'}$. Even scrambling 50% of our model's connections in this manner only yielded very small effects on performance. These experiments were described in more detail in [61] for the case of translation. In that paper we modeled Li and DiCarlo's "invariance disruption" experiments in which they showed that a temporal association paradigm can induce individual IT neurons to change their stimulus preferences under specific transformation conditions [60, 30]. We also report similar results on another "non-uniform template orbit sampling" experiment with 3D rotation-in-depth of faces in [37].

1. H. Cramer and H. Wold. Some theorems on distribution functions. J. London Math. Soc., 4:290–294, 1936.
2. J. Cuesta-Albertos, R. Fraiman, and R. T. A sharp form of the cramer–wold theorem. Journal of Theoretical Probability, 20:201–209, 2007.
3. J. Cuesta-Albertos. How many random projections suffice to determine a probability distribution? IPMs sections, 2009.
4. A. Heppes. On the determination of probability distributions of more dimensions by their projections. Acta Mathematica Hungarica, 7(3):403–410, 1956.
5. D. L. Donoho, P. B. Stark Uncertainty principles and signal recovery SIAM J. Appl. Math.,49 ,3 ,906–931 , 1989
6. D. Gabor. Theory of communication. Part 1: The analysis of information. Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of,93 ,26 ,429-441 ,1946.
7. S. Mallat. Group invariant scattering. Communications on Pure and Applied Mathematics, 65(10):1331–1398, 2012.
8. T. Poggio. On optimal nonlinear associative recall. Biological Cybernetics, 19(4):201–209, 1975.
9. T. Plate, Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations, International Joint Conference on Artificial Intelligence, 30-35, 1991.
10. T. Poggio A theory of how the brain might work Cold Spring Harb Symp Quant Biol, 1990.
11. T. Poggio, T. Vetter, and M. I. O. T. C. A. I. LAB. Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries, 1992.
12. J. Z. Leibo, J. Mutch, and T. Poggio. Why The Brain Separates Face Recognition From Object Recognition. In Advances in Neural Information Processing Systems (NIPS), Granada, Spain, 2011.
13. L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio. The timing of invariant object recognition in the human visual system. Submitted, 2013.
14. F. Anselmi, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio. Magic Materials: a theory of deep hierarchical architectures for learning sensory representations CBCL paper, Massachusetts Institute of Technology, Cambridge, MA, April 1, 2013.
15. T. Poggio and S. Smale. The mathematics of learning: Dealing with data. Notices of the American Mathematical Society (AMS), 50(5):537–544, 2003.
16. D. Arathorn. Computation in the higher visual cortices: Map-seeking circuit theory and application to machine vision. In Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop, AIPR '04, pages 73–78, Washington, DC, USA, 2004. IEEE Computer Society.
17. L. Sifre, S. Mallat, and P. France. Combined scattering for rotation invariant texture, 2012.
18. T. Poggio, S. Edelmann A network that learns to recognize three-dimensional objects. In Nature, 1990 Jan 18, 343(6255):263266.
19. J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. Computer Vision and Pattern Recognition 2006, 1:11–18, 2006.
20. H. Jhuang, E. Garrote, J. Mutch, X. Yu, V. Khilnani, T. Poggio, A. Steele and T. Serre, Automated home-cage behavioural phenotyping of mice Nature Communications, 1, 68, doi:10.1038/ncomms1064, 2010.
21. T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America, 104(15):6424–6429, 2007.
22. J.Z. Leibo, J. Mutch, L. Rosasco, S. Ullman, T. Poggio, Learning Generic Invariances in Object Recognition: Translation and Scale MIT-CSAIL-TR-2010-061, CBCL-294, 2010
23. A. Saxe, P.W. Koh, Z. Chen, M. Bhand, B. Suresh, A. Ng, On Random Weights and Unsupervised Feature Learning, Proceedings of the 28th International Conference on Machine Learning (ICML-11), 1089-1096, 2011.
24. K. Jarrett, K. Kavukcuoglu, M.A. Ranzato, Y . LeCun. What is the best multi-stage architecture for object recognition? IEEE International Conference on Computer Vision, 2146-2153, 2009.
25. N. Pinto, D. Doukhan, J.J. DiCarlo, D.D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Computational Biology, 5, 2009.
26. P. Földiák. Learning invariance from transformation sequences. Neural Computation, 3(2):194–200, 1991.
27. L. Wiskott, T.J. Sejnowski Slow feature analysis: Unsupervised learning of invariances Neural computation, 4, 14, 715-770, 2002.
28. L. Isik, J.Z. Leibo, T. Poggio Learning and disrupting invariance in visual recognition with a temporal association rule Frontiers in Computational Neuroscience,2,2012
29. N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 321(5895):1502–7, Sept. 2008.
30. N. Li and J. J. DiCarlo. Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. Neuron, 67(6):1062–1075, 2010.
31. D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex The Journal of Physiology 160, 1962.
32. M. Riesenhuber and T. Poggio. Models of object recognition. Nature Neuroscience, 3(11), 2000.
33. K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193–202, Apr. 1980.
34. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989.
35. F. Anselmi, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio. Magic Materials: a theory of deep hierarchical architectures for learning sensory representations CBCL paper, Massachusetts Institute of Technology, Cambridge, MA, April 1, 2013.
36. J. Z. Leibo, J. Mutch, and T. Poggio. Why The Brain Separates Face Recognition From Object Recognition. In Advances in Neural Information Processing Systems (NIPS), Granada, Spain, 2011.
37. Q. Liao, J.Z. Leibo, T. Poggio. Learning invariant representations and applications to face verification Advances in Neural Information Processing Systems (NIPS), 2013.
38. T. Lee and S. Soatto. Video-based descriptors for object recognition. Image and Vision Computing, 2012.
39. H. Schulz-Mirbach. Constructing invariant features by averaging techniques. In Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International. Conference on, volume 2, pages 387 –390 vol.2, 1994.
40. H. Cramer and H. Wold. Some theorems on distribution functions. J. London Math. Soc., 4:290–294, 1936.
41. A. Koloydenko. Symmetric measures via moments. Bernoulli, 14(2):362–390, 2008.
42. K. Jarrett, K. Kavukcuoglu, M.A. Ranzato, Y . LeCun. What is the best multi-stage architecture for object recognition? IEEE International Conference on Computer Vision, 2146-2153, 2009.
43. N. Pinto, D. Doukhan, J.J. DiCarlo, D.D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Computational Biology, 5, 2009.
44. O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4277–4280. IEEE, 2012.
45. A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 2012.
46. Q. V. Le, R. Monga, M. Devin, G. Corrado, K. Chen, M. Ranzato, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. CoRR,http://arxiv.org/abs/1112.6209, abs/1112.6209, 2011.
47. B.A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature, 381, 6583, 607–609, 1996.
48. S. Mallat. Group invariant scattering. Communications on Pure and Applied Mathematics, 65(10):1331–1398, 2012.
49. S. Soatto. Steps Towards a Theory of Visual Information: Active Perception, Signal-to-Symbol Conversion and the Interplay Between Sensing and Control. arXiv:1110.2053, pages 0–151, 2011.
50. S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. Mathematics of the neural response. Foundations of Computational Mathematics, 10(1):67–91, 2010.
51. T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. CBCL Paper #259/AI Memo #2005-036, 2005.
52. S. S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A Bayesian inference theory of attention. Vision Research, May 2010.
53. D. George and J. Hawkins. A hierarchical bayesian model of invariant pattern recognition in the visual cortex. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), volume 3, pages 1812–1817, 2005.
54. S. Geman. Invariance and selectivity in the ventral visual pathway. Journal of Physiology-Paris, 100(4):212–224, 2006.
55. W.S. McCulloch, W. Pitts A logical calculus of the ideas immanent in the nervous activity Bull. Math. Biophysics 5, 5115-133, 1943.
56. E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A, 2(2):284–299, 1985.
57. P. Földiák. Learning invariance from transformation sequences. Neural Computation, 3(2):194–200, 1991.
58. G. Wallis and H. H. Bülthoff. Effects of temporal association on recognition memory. Proceedings of the National Academy of Sciences of the United States of America, 98(8):4800–4, Apr. 2001.
59. D. Cox, P. Meier, N. Oertelt, and J. DiCarlo. 'Breaking'position-invariant object recognition. Nature Neuroscience, 8(9):1145–1147, 2005.
60. N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 321(5895):1502–7, Sept. 2008.
61. L. Isik, J. Z. Leibo, and T. Poggio. Learning and disrupting invariance in visual recognition with a temporal association rule. Frontiers in Computational Neuroscience, 6, 2012.
62. N. Kanwisher, Functional specificity in the human brain: a window into the functional architecture of the mind, Proceedings of the National Academy of Sciences, 107, 25, 11163, 2010.
63. D.Y. Tsao, W.A. Freiwald, Faces and objects in macaque cerebral cortex Nature, 9, 6, 989-995, 2003.
64. J.Z. Leibo, F. Anselmi, J. Mutch, A.F. Ebihara, W. Freiwald, T. Poggio, Viewinvariance and mirror-symmetric tuning in a model of the macaque face-processing system Computational and Systems Neuroscience, I-54, 2013
65. D. Marr, T. Poggio From understanding computation to understanding neural circuitry AIM-357, 1976.
66. J. Mutch and D. Lowe. Multiclass object recognition with sparse, localized features. Computer Vision and Pattern Recognition 2006, 1:11–18, 2006.
67. T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America, 104(15):6424–6429, 2007.

68. **Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, pages 255–258, 1995.**

69. **Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–97. IEEE, 2004.**

70. **C. F. Cadieu, H. Hong, D. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo. The neural representation benchmark and its evaluation on brain and machine. arXiv preprint arXiv:1301.3530, 2013.**