bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Center for Brains, Minds & Machines

CBMM Memo No. 004 **April 23, 2014**

The invariance hypothesis implies domain-specific regions in visual cortex

by

Joel Z. Leibo, Qianli Liao, Fabio Anselmi, & Tomaso Poggio

Abstract: Is visual cortex made up of general-purpose information processing machinery, or does it consist of a collection of specialized modules? If prior knowledge, acquired from learning a set of objects is only transferable to new objects that share properties with the old, then the recognition system's optimal organization must be one containing specialized modules for different object classes. Our analysis starts from a premise we call the invariance hypothesis: that the computational goal of the ventral stream is to compute an invariant-to-transformations and discriminative signature for recognition. The key condition enabling approximate transfer of invariance without sacrificing discriminability turns out to be that the learned and novel objects transform similarly. This implies that the optimal recognition system must contain subsystems trained only with data from similarly-transforming objects and suggests a novel interpretation of domain-specific regions like the fusiform face area (FFA). Furthermore, we can define an index of transformation-compatibility, computable from videos, that can be combined with information about the statistics of natural vision to yield predictions for which object categories ought to have domain-specific regions. The result is a unifying account linking the large literature on view-based recognition with the wealth of experimental evidence concerning domain-specific regions.

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

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The invariance hypothesis implies domain-specific regions in visual cortex

Joel Z. Leibo*[1], Qianli Liao[1], Fabio Anselmi[1,2], & Tomaso Poggio[1,2]

[1]Center for Brains, Minds, and Machines and the McGovern Institute for Brain Research, at the Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Istituto Italiano di Tecnologia, Genova, 16163, Italy

*To whom correspondence should be addressed: jzleibo@mit.edu

Is visual cortex made up of general-purpose information processing machinery, or does it consist of a collection of specialized modules? If prior knowledge, acquired from learning a set of objects is only transferable to new objects that share properties with the old, then the recognition system's optimal organization must be one containing specialized modules for different object classes. Our analysis starts from a premise we call the invariance hypothesis: that the computational goal of the ventral stream is to compute an invariant-to-transformations and discriminative signature for recognition. The key condition enabling approximate transfer of invariance without sacrificing discriminability turns out to be that the learned and novel objects transform similarly. This implies that the optimal recognition system must contain subsystems trained only with data from similarly-transforming objects and suggests a novel interpretation of domain-specific regions like the fusiform face area (FFA). Furthermore, we can define an index of transformation-compatibility, computable from videos, that can be combined with information about the statistics of natural vision to yield predictions for which object categories ought to have domain-specific regions. The result is a unifying account linking the large literature on view-based recognition with the wealth of experimental evidence concerning domain-specific regions.

Introduction

How can past visual experience be leveraged to improve future recognition of novel objects? Is any past experience useful for improving at-a-glance recognition of any new object? Or perhaps past experience only transfers to similar objects? Could it even be possible that past experience with certain objects actually impedes the recognition of others?[1] The answers to these questions have implications for the optimal organization of the ventral visual pathway. If prior experience is always transferable then generic circuitry would suffice. But, if any kind of similarity is a precondition for leveraging past object experience then its optimal organization would have to be a collection of specialized modules.

Consider an object recognition system trained using images of a certain set of objects A. At test time, it must recognize novel objects from a disjoint set B. Suppose that A is the union of two disjoint subsets A_{good} and A_{bad}. When training is restricted to A_{good}, test accuracy on B will be good. When restricted to A_{bad}, accuracy on the B-test will be bad. Now suppose that the test accuracy achieved by training on all of A is the

[1]If these questions sound strange it may be helpful to consider their analogs in other modalities: To decode Italian phrases, would it be more helpful to have previous experience with Spanish or with Chinese? Do prior expectations from terrestrial life impede a new astronaut's initial mobility in zero gravity?

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the average of the two. Counterintuitively, this would be a case where a system trained with more examples does worse than one trained with less. If the good and bad subsets could somehow be identified beforehand, it would be advantageous to train separate subsystems for each, and at test time, only read out results from the more successful subsystem. The remainder of this paper is concerned first with demonstrating that biological recognition problems have properties similar to the ones assumed for this toy example. Second, it will be shown there is a way that, before seeing any test data, the brain could identify which subsets of the training data ought to be separated from one another. Finally, the last part of the paper examines predictions for specific object classes. These results imply an explanation for why there are domain-specific regions like the FFA [34, 82, 9, 13] in the ventral stream.

The *invariance hypothesis* holds that the computational goal of the ventral stream is to compute from an image I a signature $\mu(I)$ (a feature vector) that is unique to each object and invariant to identity-preserving transformations like translations and rotations. It should also be invariant to non-rigid transformations that may only occur for certain object classes like the smiling of a face or the melting of an ice cube. To make this precise, let g_θ denote a transformation with parameter θ. Two images I, I' depict the same object whenever $\exists\theta$, such that $I' = g_\theta I$. Then

$$|\mu(g_\theta I) - \mu(I)| \leq \epsilon. \tag{1}$$

We say that a signature for which (1) is satisfied (for all θ) is ϵ-invariant to the family of transformations $\{g_\theta\}$. An ϵ-invariant signature that is unique to an object can be used to discriminate images of that object from images of other objects. Note that the invariance hypothesis implies that the ventral stream's goal is to provide such a signature for all objects, even those that it has not yet encountered[2].

Results

Theory sketch

One approach to modeling the ventral stream, first taken by Fukushima's Neocognitron [19], and followed by many other models [51, 60, 53, 65, 56], is based on iterating a basic module inspired by Hubel and Wiesel's proposal for the connectivity of V1 simple (AND-like) and complex (OR-like) cells. In the case of HMAX [60], each "HW"-module consists of one C-unit (corresponding to a complex cell) and all its afferent S-units (corresponding to simple cells); see fig. 1-B. The response of an S-unit to an image I is typically modeled by a dot product with a stored template t, indicated here by $\langle I, t\rangle$. Since $\langle I, t\rangle$ is maximal when $I = t$ (assuming normalized I and t), we can think of an S-unit's response as a measure of I's similarity to t. The module corresponding to Hubel and Wiesel's original proposal had several S-units, each detecting their stored template at a different position. Let $g_{\vec{x}}$ be the translation operator, when applied to an image, it returns its translation by $\vec{x}$. This lets us write the response of the specific S-unit which signals the presence of template t at position $\vec{x}$ as $\langle I, g_{\vec{x}} t\rangle$. Then, introducing a nonlinear *pooling function*, which for HMAX is usually the $\texttt{max}$ function, the response $C(I)$ of the C-unit (equivalently: the output of the HW-module, one element of the signature) is given by

$$C(I) = \texttt{max}_i \left(\langle I, g_{\vec{x}_i} t\rangle\right) \tag{2}$$

where the $\texttt{max}$ is taken over all the S-units in the module. The region of space covered by a module's S-units is called its *pooling domain* and the C-unit is said to pool the responses of its afferent S-units. More recent models based on this approach typically also pool over a range of scales [53, 65, 56]. In most cases, the first layer pooling domains are small intervals of translation and scaling. In the highest layers the pooling domains are usually global, i.e. over the entire range of translation and scaling. Notice also that the HW-module formulation is more general than HMAX. It applies to a wide class of hierarchical models of cortical

[2]This formulation of the invariance hypothesis is of independent interest beyond the argument presented here. [58, 40] describe how it can be used to derive receptive field properties including Gabor-like tuning in V1 and mirror symmetric orientation tuning curves in the anterior lateral face patch [18]. Our specific formulation is a product of Anselmi et al.'s theory of invariant recognition architectures [1], but it is also in line with other recent perspectives on the ventral stream and the recognition problem e.g., [72, 11].

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

computation, e.g., [19, 38, 61, 56]. For instance, t need not be directly interpretable as a template depicting an image of a certain object. A convolutional neural network in the sense of [39, 36] is obtained by choosing t to be the outcome of a gradient descent-based optimization procedure. In what follows we use the HW-module language since it is convenient for stating the domain-specificity argument. It is a technical exercise to translate it into the language associated with equivalent models like convolutional neural networks.

HW-modules can compute ϵ-invariant representations for a broad class of transformations [1]. However, and this is a key fact: the conditions that must be met are different for different transformations. Following Anselmi et al. [1], we can distinguish two "regimes". The first regime applies to the important special case of transformations with a group structure, e.g., 2D affine transformations. The second regime applies more broadly to any locally-affine transformation.

For a family of transformations $\{g_\theta\}$, define the *orbit* of an image I to be the set $O_I = \{g_\theta I,\ \theta \in \mathbb{R}\}$. Anselmi et al. [1] proved that HW-modules can pool over other transformations besides translation and scaling. It is possible to pool over any transformation for which orbits of template objects are available. A biologically-plausible way to learn the pooling connections within an HW-module could be to associate temporally adjacent frames of the video of visual experience (as in e.g., [17, 89, 71, 70, 28, 88]). In both regimes, the following condition is required for the invariance obtained from the orbits of a set of template objects to generalize to new objects. For all $g_\theta I \in O_I$ there is a corresponding $g_{\theta'} t \in O_t$ such that

$$\left\langle g_\theta I, t^k \right\rangle = \left\langle I, g_{\theta'} t^k \right\rangle \tag{3}$$

In the first regime, eq. (3) holds regardless of the level of similarity between the templates and test objects. Almost any templates can be used to recognize any other images invariantly to group transformations (see SI section 1). Note also that this is consistent with reports in the literature of strong performance achieved using random filters in convolutional neural networks [31, 41, 62]. Figure 1-A illustrates that the orbit with respect to in-plane rotation is invariant. The experiment shown in figure 2 verifies that for a group transformation, translation in this case, the templates need not resemble the test images. HW-modules were trained on images of translating random dot patterns and tested on faces, and vice versa. The outcome was invariant in both cases.

In the second regime, corresponding to non-group transformations, it is not possible to achieve a perfectly invariant representation. These transformations often depend on information that is not available in a single image. For example, rotation in depth depends on an object's 3D structure and illumination changes depend on its material properties (see SI section 2). Despite this, ϵ-invariance to smooth non-group transformations can still be achieved using prior knowledge of how similar objects transform. Second-regime transformations are *class-specific*, e.g., the transformation of object appearance caused by a rotation in depth is not the same 2D transformation for two objects with different 3D structures. However, by restricting to a class where all the objects have similar 3D structure, all objects do rotate (approximately) the same way. Moreover, this commonality can be exploited to transfer the invariance learned from experience with (orbits of) template objects to novel objects seen only from a single example view.

Class-specific transformations are the primary obstacle to leveraging past visual experience for future recognition of novel individuals. The simulation in figure 3 shows that HW-modules tuned to templates from the same class as the (always novel) test objects provide a signature that tolerates substantial viewpoint changes (plots on the diagonal), it also shows the deleterious effect of using templates from the wrong class (plots off the diagonal—compare to figure 2). There are many other class-specific transformations besides depth-rotation, see the supplementary information for additional simulations with illumination and body pose transformations.

How can object experience—i.e., training data—be optimally assigned to subsystems in order to maximize invariant recognition accuracy on a large and unknown set of tasks? The question is understood by considering its two extreme solutions. On the one hand, it could be that the specific objects do not matter, experience with any objects can always be leveraged for new objects. On the other hand, if the optimal assignment creates a separate group for each individual object then experience with one object will never be transferable. It

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is sufficient to consider fine-grained (or subordinate-level) recognition tasks like recognizing individual faces or particular breeds of dogs since the alternative, basic-level categorization, could be accomplished using features that do not change much under transformations (e.g., non-accidental properties [5, 3]). Furthermore, it is sufficient to consider tasks where there is only a single example image available of the test object since they are the most difficult.

More concretely, we will consider the assignment of trained HW-modules to subsystems. The response to a test image I of an HW-module trained on an orbit O_{t^k} is computed by pooling all the $\langle I, gt^k \rangle$ for all $gt^k \in O_{t^k}$. Each subsystem will consist of a number of HW-modules. For a test image I, the k-th element of the signature vector is the output of the k-th HW-module in the subsystem $\mu(I)_k = \mathtt{max}_{gt^k \in O_{t^k}} \left(\langle I, gt^k \rangle \right)$. To test a subsystem on a recognition task, compute its signature vector for each image in the set comprising the task. To get the accuracy score, compare the angle between signature vectors of images depicting same and different objects (see S.I. 5.2) for details).

Given a set of objects sampled from a category, what determines when HW-modules encoding templates for a few members of the class can be used to ϵ-invariantly recognize unfamiliar members of the category from a single example view? Recall that the transfer of ϵ-invariance depends on the condition given by eq. (3). For non-group transformations this turns out to require that the objects "transform the same way" (see SI section 1 for the proof; the notion of a "nice class" is also related [85, 42]). Given a set of orbits of different objects (only the image sequences are needed), we can then compute an index $\overline{\psi}$—which we call the *transformation compatibility*—see SI-8. $\overline{\psi}$ measures how similarly the objects in the class transform. If an object category has too low $\overline{\psi}$, then there would be no gain from creating a subsystem for that category. Whenever a category has high $\overline{\psi}$, it is a candidate for having a dedicated subsystem.

For the special case of rotation in depth, we can use 3D modelling / rendering software [6] to obtain (dense samples from) the orbits of any objects for which we can obtain 3D models. We computed the transformation compatibility index $\overline{\psi}$ for several datasets from different sources (see methods). Faces had the highest $\overline{\psi}$ of any naturalistic category we tested—unsurprising since recognizability likely influenced face evolution. A set of chair objects (from [12]) had very low $\overline{\psi}$ implying no benefit would be obtained from a chair-specific region. More interestingly, we tested a set of synthetic "wire" objects, very similar to those used in many classic experiments on view-based recognition e.g. [7, 44, 45]. We found that the wire objects had the lowest $\overline{\psi}$ of any category we tested; experience with familiar wire objects does not transfer to new wire objects. It is never productive to group them into a subsystem.

Simulations

The analysis so far has been on the computational/algorithmic level [48]. It answers questions about which categories could or could not have productive subsystems. To arrive at predictions for which domain-specific regions will be found in the brain, we adopt the hypothesis that the neural circuitry implementing a subsystem must be localized on cortex. There are several ways to justify this hypothesis (see the discussion section below).

Any model that can predict which specific categories will have domain-specific regions must depend on contingent facts about the world, in particular, the—difficult to approximate—distribution $\mathcal{D}$ of objects and their transformations encountered during natural vision. Nevertheless, the present proposal does suggest a family of models of ventral stream development yielding specific predictions (see also SI-4). Consider the following: HW-modules may be assigned to cluster near one another on cortex in order to maximize the transformation compatibility $\overline{\psi}$ of the set of objects represented in each local neighbourhood. Whenever a new object is learned, its HW-module could be placed on cortex in the neighbourhood with which it transforms most compatibly. We conjecture that if this iterative clustering process were simulated, at each iteration sampling a new object from $\mathcal{D}$, then, the resulting cortex model obtained after some time would have a small number of very large clusters, probably corresponding to faces, bodies, and orthography in a literate brain's native language. The rest of the objects would be encoded by HW-modules at random locations. Since neuroimaging

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methods like fMRI have limited resolution, only the largest clusters would be visible to them. Cortical regions with low $\overline{\psi}$ would appear in neuroimaging experiments as generic "object regions" like LOC [46].

Since we cannot sample from $\mathcal{D}$, we cannot actually perform the simulation outlined in the previous paragraph. However, by assuming particular distributions and sampling from a library of $\sim$10,000 3D models [12, 68], we can study the special case where the only transformation is rotation in depth. Figure 5 shows example clusters obtained this way. A key result of our analysis is that the development of domain specific regions depends not only on object frequencies but also on transformation compatibility. Thus one way of arguing could be to show that in some cases, the predictions could be robust over a reasonably large range of statistical assumptions. For example, we found across three experiments, each using a different object distribution, that robust face and body clusters always appeared (SI-4). Due to the strong effect of $\overline{\psi}$, a face cluster formed even when the distribution of objects was biased *against* faces as in figure 6.

While a view-invariant basic-level categorization task, cars vs. airplanes, can be performed to similar accuracy using any of the clusters (figure 6-C), performance on the analogous view-invariant face verification task was significantly higher when the face cluster was used (figure 6-B). This illustrates that $\overline{\psi}$-based clustering into subsystems is only beneficial for the subordinate level task.

Discussion

Why are there domain-specific regions in the anterior ventral stream but not the posterior ventral stream [81, 37]? The templates used to implement invariance to group transformations need not be changed for different object classes while the templates implementing non-group invariance are class-specific. Thus it is efficient to put the generic circuitry of the first regime in the hierarchy's early stages, postponing the need to branch to different domain-specific regions tuned to specific object classes until later, i.e., more anterior, stages. Recent studies of the macaque face-processing system [18, 30] show that category selectivity develops in a series of steps, with posterior face regions less face selective than anterior ones. Additionally, there is a progression from a view-specific face representation in earlier regions to a view-tolerant representation in the most anterior region [18]. Both findings could be accounted for in a face-specific hierarchical model that increases in template size and pooling region size with each subsequent layer (e.g., [23, 84]). The use of large face-specific templates may be an effective way to gate the entrance to the face-specific subsystem so as to keep out spurious activations from non-faces. The algorithmic effect of large face-specific templates is to confer tolerance to clutter [43]. These results are particularly interesting in light of models showing that large face templates are sufficient to explain holistic effects observed in psychophysics experiments [90, 73].

Why should the circuitry comprising a subsystem be localized on cortex? In principle, any HW-module could be anywhere, as long as the wiring all went to the right place. However, there are several reasons to think that the actual constraints under which the brain operates and its available information processing mechanisms favor a situation in which, at each level of the hierarchy, all the specialized circuitry for one domain is in a localized region of cortex, separate from the circuitry for other domains. In particular, wiring length considerations are likely to play a role here [59, 4, 52, 8]. Another possibility is that cortical localization enables the use of neuromodulatory mechanisms that act on local neighborhoods of cortex to affect all the circuitry for a particular domain at once [47]. Attention may operate through mechanisms of this sort—possibly involving acetylcholine or norepinephrine [91].

There are other domain-specific regions in the ventral stream besides faces and bodies; we consider several of them in light of our results here. It is possible that even more regions for less-common (or less transformation-compatible) object classes would appear with higher resolution scans. One example may be the fruit area, discovered in macaques with high-field fMRI [37].

1. **Lateral Occipital Complex (LOC) [46]**
   These results imply that LOC is not really a dedicated region for general object processing. Rather, it is a heterogeneous area of cortex containing many domain-specific regions too small to be detected with the resolution of fMRI. It may also include clusters that are not dominated by one object category as we

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sometimes observed appearing in simulations (see S.I. section 4).

2. **Parahippocampal Place Area (PPA) [15]**
   A ventral stream region that appears to be specialized for scene processing seems, at first, to be problematic for our hypothesis. It is unclear whether or not there are any transformations with respect to which the category of "scene" would be compatible. One possibility, which we considered in preliminary work, is the hypothesis that "perspective", i.e., depth-cues from 2D images could be a transformation with this property [35]. Another possibility could be that the PPA is not really a ventral stream domain-specific region in the same sense as the Fusiform Face Area (FFA) or the Extrastriate Body Area (EBA). After all, it is arguable that it is not really properly considered part of the ventral stream. In particular, Schacter, Bar, and others in the medial temporal lobe memory literature, have emphasized parahippocampal cortex's role in contextual associations and constructive simulation of future events over place/scene processing [2, 63].

3. **The Visual Word Form Area (VWFA) [9]**
   In addition to the generic transformations that apply to all objects, printed words undergo several non-generic transformations that never occur with other objects. We can read despite the large image changes occurring when a page is viewed from a different angle. Additionally, many properties of printed letters change with typeface, but our ability to read—even in novel fonts—is preserved. Reading hand-written text poses an even more severe version of the same computational problem. Thus, VWFA is well-accounted for by the invariance hypothesis. Words are frequently-viewed stimuli which undergo class-specific transformations.

What about the alternative hypothesis: that we have domain-specific regions for faces, bodies, etc, just because these classes are important? The nativist version of this question is answered by the existence of the VWFA [9]. The empiricist version is that we develop domain-specific regions for any objects we see frequently enough, or need to perform certain tasks on. The expertise hypothesis is a sophisticated example [20, 77, 10, 87]. The implications of our invariance hypothesis are at odds with some purely expertise-based accounts of domain-specificity. The two primary points of contention are 1. we propose that transformation compatibility is the critical factor driving the development of domain-specific regions, and 2. the invariance hypothesis implies a separation of the circuitry for object classes that transform differently from one another. Thus, unless the greeble objects of [20, 77] happen to transform similarly to faces (an unlikely event), our model would predict that the circuitry underlying expert-level greeble recognition would have to be separate from that underlying face recognition. It is worth noting that studies claiming to have demonstrated expertise-related selectivity of FFA have also been criticized on empirical and methodological grounds, cf. [33]. In any case, the debate could be resolved by an experiment along the lines of Freiwald and Tsao's showing the face-specific hierarchy computes a 3D rotation-tolerant representation [18]. The same methods could be employed with animals trained for visual expertise in a non-face category. Importantly, the predictions of the present proposal would depend on the transformation compatibility of the chosen category.

Is this proposal at odds with the literature emphasizing the view-dependence of human vision when tested on subordinate level tasks with unfamiliar examples—e.g.[7, 80, 75]? We believe it is consistent with most of this literature. We merely emphasize the substantial view-*tolerance* achieved for certain object classes, while they emphasize the lack of complete invariance. Their emphasis was appropriate in the context of earlier debates about view-invariance [49, 5, 83, 57], and before differences between the view-tolerance achieved on basic-level and subordinate-level tasks were fully appreciated [74, 64, 76].

The view-dependence observed in experiments with novel faces [80, 24] is consistent with the predictions of our theory. The 3D structure of faces does not vary wildly within the class, but there is still some significant variation. It is this variability in 3D structure within the class that is the source of the imperfect performance in our simulations. Many psychophysical experiments on viewpoint invariance were performed with synthetic "wire" objects defined entirely by their 3D structure e.g., [7, 44, 45]. We found that they were by far, the least transformation-compatible (lowest $\overline{\psi}$) objects we tested (fig. 4). Thus our proposal predicts particularly weak

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

performance on viewpoint-tolerance tasks with novel examples of these stimuli and that is precisely what is observed [44].

Tarr and Gauthier (1998) found that learned viewpoint-dependent mechanisms could generalize across members of a homogenous object class [76]. They tested both homogenous block-like objects, and several other classes of more complex novel shapes. They concluded that this kind of generalization was restricted to visually similar objects. These results seem to be consistent with our proposal. Additionally, our hypothesis predicts better within-class generalization for object classes with higher $\overline{\psi}$. That is, transformation compatibility, not visual similarity per se, may be the factor influencing the extent of within-class generalization of learned view-tolerance.

An alternative account holds that visual representations are distributed across the entire ventral visual pathway [27, 22]. Such proposals are motivated by the fact that it is often possible to categorize stimuli on the basic level using the pattern of BOLD response they elicit over all of occipito-temporal cortex, even when category-selective regions are excluded from the analysis. However, analogous experiments with subordinate level tasks have shown that BOLD responses of category-selective regions cannot distinguish between pairs of non-preferred categories [69], and single unit activity patterns in the macaque anterior medial face-selective patch can identify individuals [18]. Moreover, such distributed representations are implausible in light of neuropsychological studies of acquired prosopagnosia patients [86] and the recent demonstration that stimulating cells in FFA distorts perception of faces but not other objects [55].

How should these results be understood in light of recent reports of very strong performance of computer vision systems employing apparently generic circuitry for object recognition tasks e.g., [36, 92]? We note that these advances, while exciting, address basic-level categorization, a different problem from the one considered here. Class-specific transformation invariance is mainly only an issue when discriminating individuals among very similar distractors. The interference arising from training across objects that do not transform compatibly only arises in the multi-category fine-grained setting where a single system is met with multiple different fine-grained (or subordinate level) tasks.

But, what about the old AI dream of a universal cortical algorithm that would work on all types of data? Do these results mean we need to give up on those ideas? Must we believe the brain is just a bag of hacks for different circumstances that we can never hope to understand by applying general principles? We do not think such pessimism is the correct reaction. The proposed theory suggests that all domain-specific regions perform the same basic computations, merely using different subsets of the training data[3]. It remains possible that the entire class-specific hierarchy could develop through the operation of a single learning rule. The problem of determining at each moment which subsystem should be in control is easily dealt with by a highly biologically-plausible gating mechanism that simply ignores responses below a certain threshold. One can view the project of giving a computational-level explanation for the organization of the anterior ventral stream as a rescue operation, reconciling the hopes of those building general-purpose learning systems with the fact of domain-specific regions in visual cortex.

[3]The theory predicts a "modularity of content" as opposed to "modularity of process" cf. [16].

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

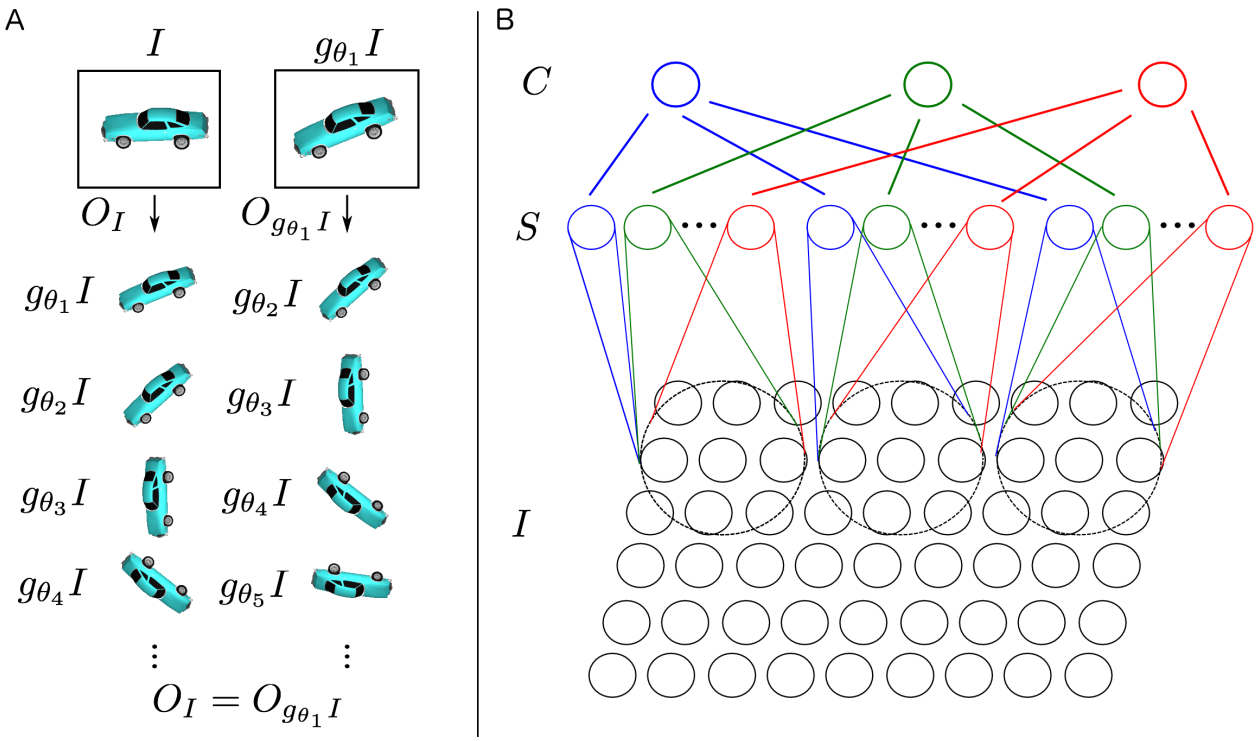

Figure 1: **A.** Illustration that the orbit with respect to in-plane rotation is invariant and unique. **B.** Three HW-modules are shown. In this example, each HW-module pools over a 9×3 region of the image. Each S-unit stores a 3×3 template and there are three S-units per HW-module.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

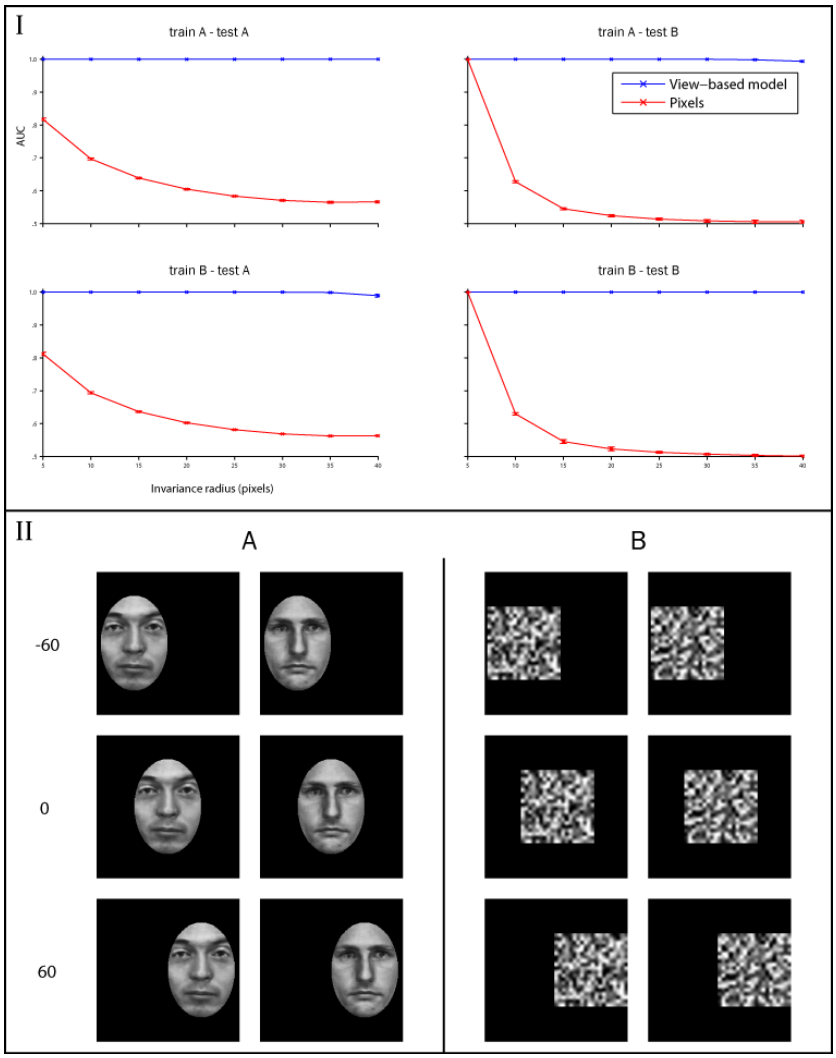

Figure 2: Translation invariance. Bottom panel (II): Example images from the two classes. The faces were obtained from the Max-Planck Institute dataset [80] and then contrast normalized and translated over a black background. Top panel (I): The left column shows the results of a test of translation invariance for faces and the right column shows the same test for random noise patterns. The view-based model (blue curve) was built using templates from class A in the top row and class B in the bottom row. The abscissa of each plot shows the maximum invariance range (a distance in pixels) over which target and distractor images were presented. The view-based model was never tested on any of the images that were used as templates. Error bars (+/- one standard deviation) were computed over 5 cross validation runs using different (always independent) template and testing images.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

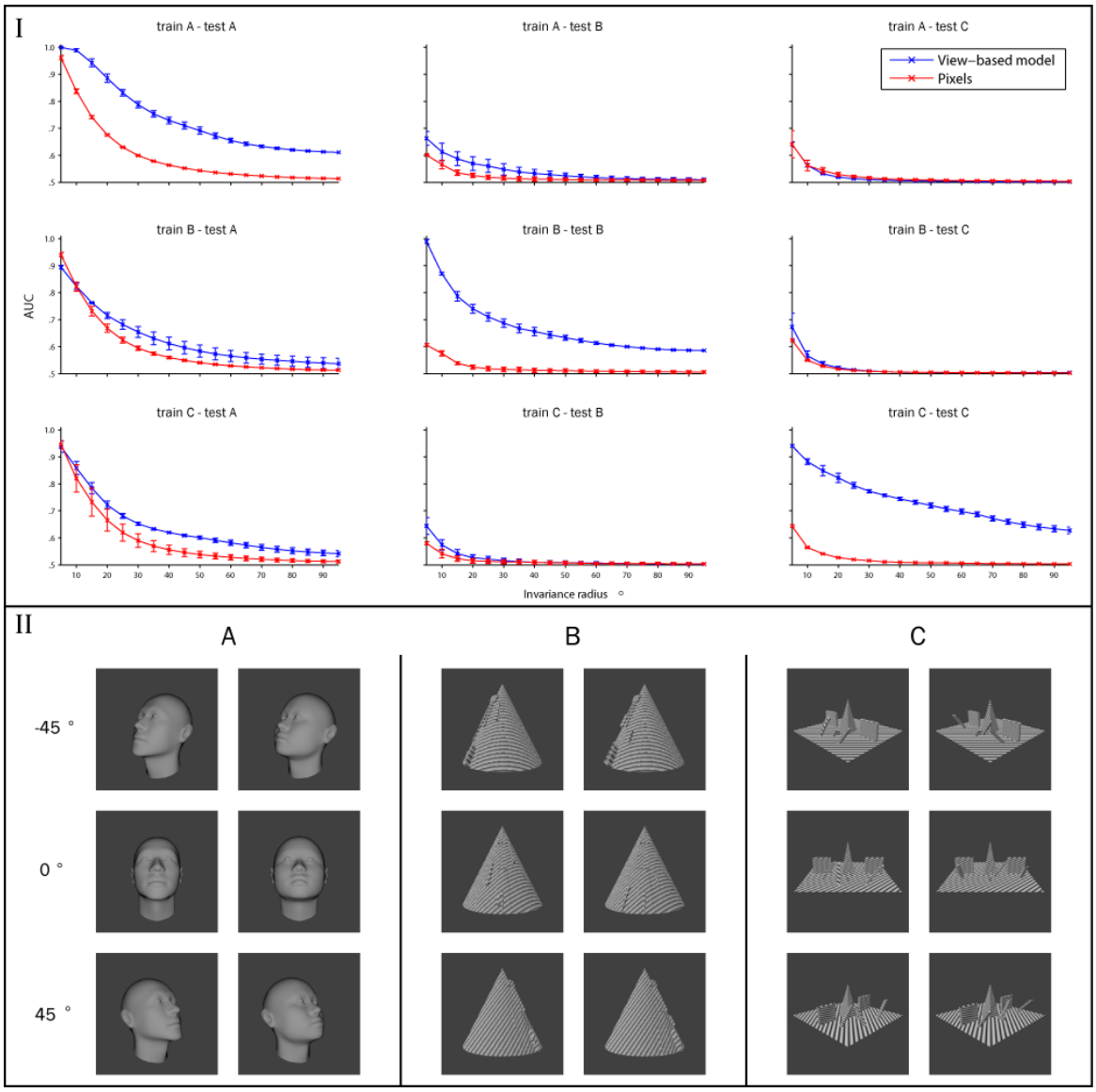

Figure 3: Class-specific transfer of depth-rotation invariance. Bottom panel (II): Example images from the three classes. Top panel (I): The left column shows the results of a test of 3D rotation invariance on faces (class A), the middle column shows results for class B and the right column shows the results for class C. The view-based model (blue curve) was built using images from class A in the top row, class B in the middle row, and class C in the bottom row. The abscissa of each plot shows the maximum invariance range (degrees of rotation away from the frontal face) over which target and distractor images were presented. The view-based model was never tested on any of the images that were used as templates. Error bars (+/- one standard deviation) were computed over 20 cross validation runs using different choices of template and test images. Only the plots on the diagonal (train A - test A, train B - test B, train C- test C) show an improvement of the view-based model over the pixel representation. That is, only when the test images transform similarly to the templates is there any benefit from pooling.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Object class	Transformation	$\overline{\psi}$
Chairs	Rotation in depth	0.00540
Fig. 3 faces	Rotation in depth	0.57600
Fig. 3 class B	Rotation in depth	0.95310
Fig. 3 class C	Rotation in depth	0.83800
Fig. 3 all classes	Rotation in depth	0.26520
COIL-100 [54]	Rotation in depth	0.00630
Wire objects [44]	Rotation in depth	-0.00007

Figure 4: Table of transformation compatibilities. COIL-100 is a library of images of 100 common household items photographed from a range of orientations using a turntable [54]. The wire objects resemble those used in psychophysics and physiology experiments: [7, 44, 45]. They were generated by following the same protocol as used in those studies.

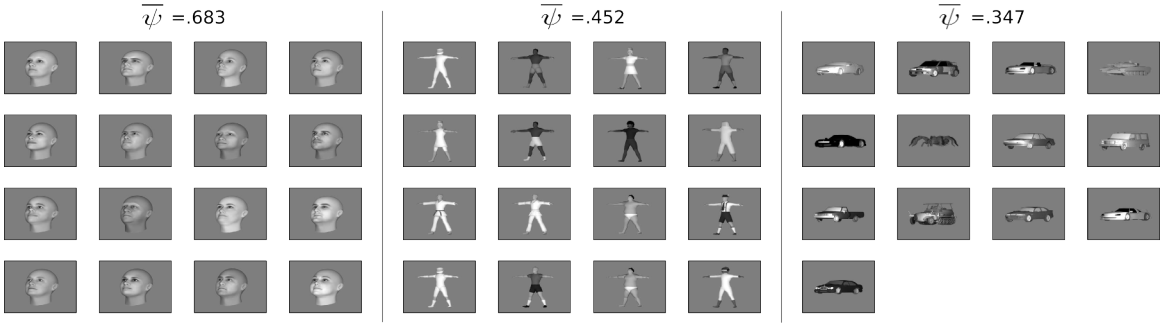

Figure 5: Three example clusters that developed in a simulation with an object distribution biased against faces (the same simulation as in figures S13-C, S14-C, and S15-C).

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

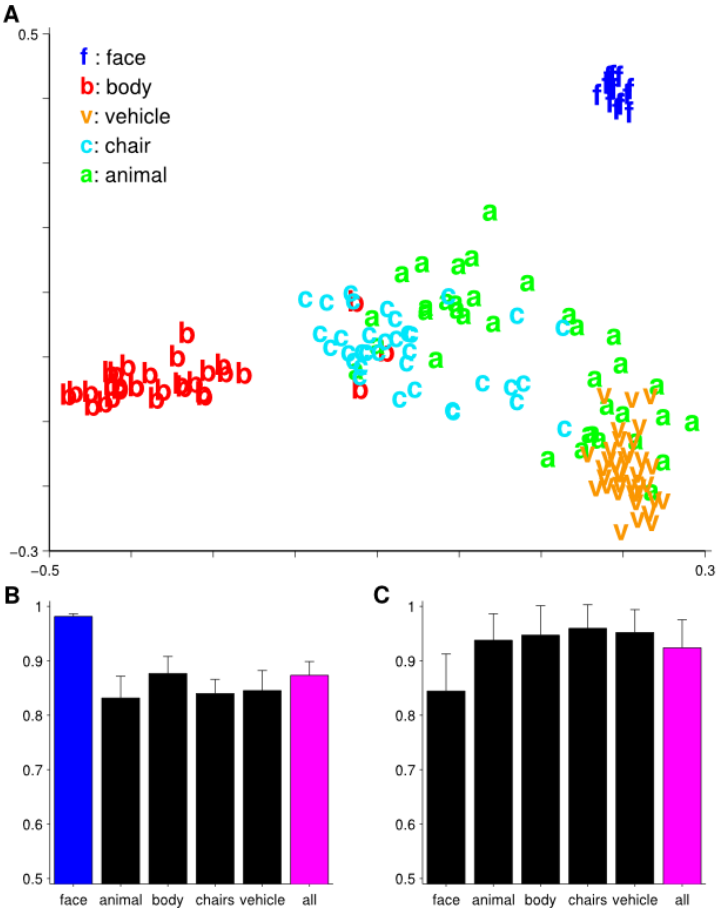

Figure 6: Simulation of the development of domain-specific regions. In this case the distribution of objects was biased against faces (faces were only 16 of the 156 objects in this simulation). Depth-rotation is the only transformation used here. The main assumption is that the distance along cortex between two HW-modules for two different templates is proportional to how similarly the two templates transform. See SI-4 for results of the analogous simulations using different object distributions **A.** Multidimensional scaling plot based on pairwise transformation compatibility ψ. **B.** Results on a test of view-invariant face verification (same-different matching). Each bar corresponds to a different cluster produced by an iterative clustering algorithm based on $\overline{\psi}$ which models visual development—see supplementary methods. The labels on the abscissa correspond to the dominant category in the cluster. **C.** Basic-level categorization results: Cars versus airplanes. Error bars were obtained by repeating the experiment 5 times, presenting the objects in a different random order during development and randomly choosing different objects for the test set.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References and Notes

[1] Fabio Anselmi, Joel Z. Leibo, Jim Mutch, Lorenzo Rosasco, Andrea Tacchetti, and Tomaso Poggio. Unsupervised Learning of Invariant Representations in Hierarchical Architectures. *arXiv:1311.4158v3 [cs.CV]*, 2013.

[2] M Bar, E Aminoff, and Daniel L. Schacter. Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. *The Journal of Neuroscience*, 28(34):8539–8544, 2008. URL http://www.jneurosci.org/content/28/34/8539.short.

[3] I Bar M, Biederman. One-shot viewpoint invariance in matching novel objects. *Vision Research*, 39(17):2885–2899, August 1999. ISSN 00426989. doi: 10.1016/S0042-6989(98)00309-5. URL http://dx.doi.org/10.1016/S0042-6989(98)00309-5.

[4] HB Barlow. Why have multiple cortical areas? *Vision Research*, 26(1):81–90, 1986. URL http://www.sciencedirect.com/science/article/pii/0042698986900726.

[5] I Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. URL http://psycnet.apa.org/journals/rev/94/2/115/.

[6] Blender.org. Blender 2.6, 2013.

[7] H.H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, 89(1):60, 1992. URL http://www.pnas.org/content/89/1/60.short.

[8] Dimitri B. Chklovskii and Alexei A. Koulakov. Maps in the brain: What can we learn from them? *Annual Review of Neuroscience*, 27:369–392, 2004. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.27.070203.144226.

[9] L Cohen, S Dehaene, and L Naccache. The visual word form area. *Brain*, 123(2):291, 2000. URL http://brain.oxfordjournals.org/content/123/2/291.short.

[10] M.N. Dailey and G.W. Cottrell. Organization of face and object recognition in modular neural network models. *Neural Networks*, 12(7-8):1053–1074, 1999. URL http://linkinghub.elsevier.com/retrieve/pii/s0893-6080(99)00050-7.

[11] James J. DiCarlo, D Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73 (3):415–434, 2012. URL http://www.sciencedirect.com/science/article/pii/S089662731200092X.

[12] Digimation.com. Digimation archive.

[13] PE Downing and Y Jiang. A cortical area selective for visual processing of the human body. *Science*, 293(5539): 2470, 2001. URL http://www.sciencemag.org/content/293/5539/2470.short.

[14] PE Downing and MV Peelen. The role of occipitotemporal body-selective regions in person perception. *Cognitive Neuroscience*, 2(3-4):186–203, 2011. URL http://www.tandfonline.com/doi/abs/10.1080/17588928.2011.582945.

[15] R Epstein and N Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998. URL http://www.nature.com/nature/journal/v392/n6676/abs/392598a0.html.

[16] Jerry A Fodor. *The modularity of mind: An essay on faculty psychology*. MIT press, 1983.

[17] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991. URL http://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.2.194.

[18] Winrich A. Freiwald and D.Y. Tsao. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, 330(6005):845, 2010. ISSN 0036-8075. URL http://www.sciencemag.org/cgi/content/abstract/330/6005/845.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[19] Kunihiko 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, April 1980. ISSN 0340-1200. doi: 10.1007/BF00344251. URL http://www.springerlink.com/content/r6g5w3tt54528137.

[20] Isabel Gauthier and Michael J. Tarr. Becoming a" greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682, 1997. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.1513&rep=rep1&type=pdf.

[21] D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*. Peninsula Publishing, Los Altos, CA, USA, 1989.

[22] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425, 2001. URL http://www.sciencemag.org/content/293/5539/2425.short.

[23] Bernd Heisele, Thomas Serre, Massimiliano Pontil, and Tomaso Poggio. Component-based Face Detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 659–657, Kauai, Hawaii, USA, 2001. IEEE. doi: 10.1.1.8.957. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.957.

[24] H Hill, Phillipe G. Schyns, and S Akamatsu. Information and viewpoint dependence in face recognition. *Cognition*, 62(2):201–222, 1997. URL http://www.sciencedirect.com/science/article/pii/S0010027796007858.

[25] DH Hubel and TN Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106, 1962. URL http://jp.physoc.org/content/160/1/106.full.pdf.

[26] Chou P. Hung, Gabriel Kreiman, Tomaso Poggio, and James J. DiCarlo. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866, November 2005. doi: 10.1126/science.1117593. URL http://www.sciencemag.org/cgi/content/abstract/310/5749/863.

[27] Alumit Ishai, Leslie G Ungerleider, Alex Martin, Jennifer L Schouten, and James V Haxby. Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16):9379–9384, 1999.

[28] Leyla Isik, Joel Z. Leibo, and Tomaso Poggio. Learning and disrupting invariance in visual recognition with a temporal association rule. *Front. Comput. Neurosci.*, 6(37), 2012. doi: 10.3389/fncom.2012.00037. URL http://www.frontiersin.org/Computational_Neuroscience/10.3389/fncom.2012.00037/abstract.

[29] Leyla Isik, Ethan M. Meyers, Joel Z. Leibo, and T. Poggio. The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, 111:91–102, 2013.

[30] Elias B. Issa and James J. DiCarlo. Precedence of the Eye Region in Neural Processing of Faces. *The Journal of Neuroscience*, 32(47):16666–16682, 2012. URL http://www.jneurosci.org/content/32/47/16666.short.

[31] K. Jarrett, K. Kavukcuoglu, MA Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *IEEE International Conference on Computer Vision*, pages 2146–2153, 2009. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5459469.

[32] 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. URL http://www.pnas.org/content/107/25/11163.short.

[33] N. Kanwisher and G. Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2109, 2006. URL http://rstb.royalsocietypublishing.org/content/361/1476/2109.abstract.

[34] N. Kanwisher, J. McDermott, and M.M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302, 1997. URL http://www.jneurosci.org/content/17/11/4302.short.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[35] Emily Y. Ko, Joel Z. Leibo, and Tomaso Poggio. A hierarchical model of perspective-invariant scene identification. In *Society for Neuroscience (486.16/OO26).*, Washington D.C., 2011. URL http://cbcl.mit.edu/publications/ps/sfn_2011_perspect_poster_V1.pdf.

[36] A Krizhevsky, I Sutskever, and G Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1106–1114, Lake Tahoe, CA, 2012. URL http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf.

[37] S.P. Ku, A.S. Tolias, N.K. Logothetis, and J. Goense. fMRI of the Face-Processing Network in the Ventral Temporal Lobe of Awake and Anesthetized Macaques. *Neuron*, 70(2):352–362, 2011. URL http://linkinghub.elsevier.com/retrieve/pii/S0896627311002054.

[38] Yann LeCun, O Matan, B Boser, J. S. Denker, D Henderson, RE Howard, W Hubbard, L. D. Jacket, and H. S. Baird. Handwritten zip code recognition with multilayer networks. In *Proceedings of the 10th International Conference on Pattern Recognition*, volume 2, pages 35–40. IEEE, 1990. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=119325.

[39] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 90–97, 2004. URL http://www.computer.org/portal/web/csdl/doi/10.1109/CVPR.2004.144.

[40] Joel Z. Leibo. *The Invariance Hypothesis and the Ventral Stream*. PhD thesis, Massachusetts Institute of Technology, 2013. URL http://cbcl.csail.mit.edu/projects/cbcl/publications/theses/thesis-leibo.pdf.

[41] Joel Z. Leibo, Jim Mutch, Lorenzo Rosasco, Shimon Ullman, and Tomaso Poggio. Learning Generic Invariances in Object Recognition: Translation and Scale. *MIT-CSAIL-TR-2010-061, CBCL-294*, 2010. URL http://hdl.handle.net/1721.1/60378.

[42] Joel Z. Leibo, James Mutch, and Tomaso Poggio. Why The Brain Separates Face Recognition From Object Recognition. In *Advances in Neural Information Processing Systems (NIPS)*, Granada, Spain, 2011. URL http://cbcl.mit.edu/publications/ps/Leibo_Mutch_Poggio_face_invar_v08_cam_rdy_letter_Dec2011.pdf.

[43] Q Liao, JZ Leibo, Y Mroueh, and T Poggio. Can a biologically-plausible hierarchy effectively replace face detection, alignment, and recognition pipelines? *arXiv preprint arXiv:1311.4082*, 2014. URL http://arxiv.org/abs/1311.4082.

[44] NK Logothetis, J. Pauls, HH Bülthoff, and T. Poggio. View-dependent object recognition by monkeys. *Current Biology*, 4(5):401–414, 1994. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982200000890.

[45] NK Logothetis, J Pauls, and T Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563, 1995. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982295001084.

[46] R. Malach, J. B. Reppas, R. R. Benson, K.K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–8139, 1995. URL http://www.pnas.org/content/92/18/8135.short.

[47] E Marder. Neuromodulation of neuronal circuits: back to the future. *Neuron*, 76(1):1–11, 2012. URL http://www.sciencedirect.com/science/article/pii/S0896627312008173.

[48] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc., New York, NY, 1982. URL http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=12242.

[49] D Marr and Herbert Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294, 1978. URL http://rspb.royalsocietypublishing.org/content/200/1140/269.short.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[50] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. URL http://link.springer.com/article/10.1007/BF02478259.

[51] Bartlett W. Mel. SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition. *Neural Computation*, 9(4):777–804, May 1997. doi: 10.1162/neco.1997.9.4.777. URL http://dx.doi.org/10.1162/neco.1997.9.4.777 http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.4.777.

[52] G Mitchison. Neuronal branching patterns and the economy of cortical wiring. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 245(1313):151–158, 1991. URL http://rspb.royalsocietypublishing.org/content/245/1313/151.short.

[53] J Mutch and DG Lowe. Multiclass object recognition with sparse, localized features. *Computer Vision and Pattern Recognition 2006*, 1:11–18, 2006. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1640736.

[54] SA Nene, SK Nayar, and H Murase. Columbia Object Image Library (COIL-100). *Columbia University Tech. Report No. CUCS-006-96*, 1996.

[55] Josef Parvizi, Corentin Jacques, Brett L Foster, Nathan Withoft, Vinitha Rangarajan, Kevin S Weiner, and Kalanit Grill-Spector. Electrical stimulation of human fusiform face-selective regions distorts face perception. *The Journal of Neuroscience*, 32(43):14915–14920, 2012.

[56] N Pinto, D Doukhan, JJ DiCarlo, and DD Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 5(11), 2009. URL http://dx.plos.org/10.1371/journal.pcbi.1000579.

[57] T Poggio and S Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266, 1990. URL http://cbcl.mit.edu/people/poggio-new/journals/poggio-edelman-nature-1990.pdf.

[58] Tomaso Poggio, Jim Mutch, Fabio Anselmi, Andrea Tacchetti, Lorenzo Rosasco, and Joel Z. Leibo. Does invariant recognition predict tuning of neurons in sensory cortex? *MIT-CSAIL-TR-2013-019, CBCL-313*, 2013.

[59] Santiago Ramon y Cajal. *Texture of the Nervous System of Man and the Vertebrates: I*. Springer, 1999.

[60] M Riesenhuber and T Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, November 1999. ISSN 1097-6256. doi: 10.1038/14819. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.46.7843&rep=rep1&type=pdf.

[61] ET Rolls. Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Frontiers in Computational Neuroscience*, 6, 2012. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3378046/.

[62] A. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng. On random weights and unsupervised feature learning. *Proceedings of the International Conference on Machine Learning (ICML)*, 2011. URL http://ai.stanford.edu/ ang/papers/nipsdlufl10-RandomWeights.pdf.

[63] Daniel L. Schacter and Donna R. Addis. On the nature of medial temporal lobe contributions to the constructive simulation of future events. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521): 1245–1253, 2009. URL http://rstb.royalsocietypublishing.org/content/364/1521/1245.short.

[64] Phillipe G. Schyns. Diagnostic recognition: task constraints, object information, and their interactions. *Cognition*, 67 (1):147–179, 1998. URL http://www.sciencedirect.com/science/article/pii/S001002779800016X.

[65] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007. URL http://portal.acm.org/citation.cfm?id=1263421&dl=.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[66] Thomas Serre, Aude Oliva, and Tomaso 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. ISSN 0027-8424. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=18713198.

[67] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. URL http://www.annualreviews.org/doi/pdf/10.1146/annurev.neuro.24.1.1193.

[68] Singular Inversions. FaceGen Modeller 3, 2003.

[69] Mona Spiridon and Nancy Kanwisher. How distributed is visual category information in human occipito-temporal cortex? an fmri study. *Neuron*, 35(6):1157–1165, 2002.

[70] MW Spratling. Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):753–761, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1407878.

[71] S.M. Stringer and E.T. Rolls. Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation*, 14(11):2585–2596, 2002. URL http://www.mitpressjournals.org/doi/abs/10.1162/089976602760407982.

[72] G Sundaramoorthi, P Petersen, V. S. Varadarajan, and S Soatto. On the set of images modulo viewpoint and contrast changes. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 832–839, 2009. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5206704.

[73] C Tan and T Poggio. Faces as a "Model Category" for Visual Object Recognition. *MIT-CSAIL-TR-2013-004, CBCL-311*, 2013. URL http://dspace.mit.edu/handle/1721.1/77936.

[74] Michael J. Tarr and Heinrich H. Bülthoff. Is human object recognition better described by geon structural descriptions or by multiple views? *Journal of Experimental Psychology: Human Perception and Performance*, 21(6):1494–1505, 1995. URL http://psycnet.apa.org/journals/xhp/21/6/1494/.

[75] Michael J. Tarr and HH Bülthoff. Image-based object recognition in man, monkey and machine. *Cognition*, 67(1): 1–20, 1998. URL http://www.sciencedirect.com/science/article/pii/S0010027798000262.

[76] Michael J. Tarr and I Gauthier. Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition*, 67(1):73–110, 1998. URL http://www.sciencedirect.com/science/article/pii/S0010027798000237.

[77] Michael J. Tarr and Iisabel Gauthier. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, 3:764–770, 2000. URL http://psych.colorado.edu/ kimlab/Tarr.Gauthier.commentary.nn2000.pdf.

[78] S Thorpe, D Fize, and C Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996. URL http://www1.nin.knaw.nl/ korjouko/download/claire/Speed of processing in the human visual system. S.Thorpe D.Fize C.Marlot. 1996.pdf.

[79] Warren S. Torgerson. *Theory and methods of scaling.* Wiley, 1958. URL http://psycnet.apa.org/psycinfo/1959-07320-000.

[80] NF Troje and H.H. Bülthoff. Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771, 1996. URL http://linkinghub.elsevier.com/retrieve/pii/0042698995002308.

[81] D.Y. Tsao, Winrich A. Freiwald, T.A. Knutsen, J.B. Mandeville, and R.B.H. Tootell. Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995, 2003. URL http://www.nature.com/neuro/journal/v6/n9/abs/nn1111.html.

[82] D.Y. Tsao, Winrich A. Freiwald, R.B.H. Tootell, and M.S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670, 2006. URL http://www.sciencemag.org/content/311/5761/670.short.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[83] S. Ullman. Aligning pictorial descriptions: An approach to object recognitionâĂć 1. *Cognition*, 32(3):193–254, 1989. ISSN 0010-0277. URL http://linkinghub.elsevier.com/retrieve/pii/001002778990036X.

[84] S Ullman and B Epshtein. Visual classification by a hierarchy of extended fragments. In *Toward Category-Level Object Recognition*, pages 321–344. Springer, 2006. URL http://link.springer.com/chapter/10.1007/11957959_17.

[85] T. Vetter, A. Hurlbert, and T. Poggio. View-based models of 3D object recognition: invariance to imaging transformations. *Cerebral Cortex*, 5(3):261, 1995. ISSN 1047-3211. URL http://cercor.oxfordjournals.org/content/5/3/261.abstract.

[86] Y Wada and T Yamamoto. Selective impairment of facial recognition due to a haematoma restricted to the right fusiform and lateral occipital region. *Journal of Neurology, Neurosurgery & Psychiatry*, 71(2):254–257, 2001.

[87] G Wallis. Toward a unified model of face and object recognition in the human visual system. *Frontiers in psychology*, 4(497), 2013. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3744012/.

[88] Tristan J Webb and Edmund Rolls. Deformation-specific and deformation-invariant visual object recognition: pose vs. identity recognition of people and deforming objects. *Frontiers in Computational Neuroscience*, 8:37, 2014.

[89] L. Wiskott and T.J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14 (4):715–770, 2002. URL http://www.mitpressjournals.org/doi/abs/10.1162/089976602317318938.

[90] Andrew W Young, Deborah Hellawell, and Dennis C Hay. Configurational information in face perception. *Perception*, 16(6):747–759, 1987. URL http://www.perceptionweb.com/perception/fulltext/p16/p160747.pdf.

[91] AJ Yu and P Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692, 2005. URL http://www.sciencedirect.com/science/article/pii/S0896627305003624.

[92] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional neural networks. *arXiv preprint arXiv:1311.2901*, 2013.

Acknowledgments

We would like to thank users bohemix and SoylentGreen of the Blender Open Material Repository for contributing the materials used to create the images for the illumination simulations. We also thank Leyla Isik, Chris Summerfield, Winrich Freiwald, Pawan Sinha, and Nancy Kanwisher for their comments on early versions of this manuscript, and Heejung Kim for her help preparing one of the supplementary figures. This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This material is based upon work supported by the Center for Brains, Minds, 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.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Information

Contents

1 A theory of architectures for invariant object recognition **21**
1.1 The first regime: generic invariance . . . 21
1.2 The second regime: class-specific invariance . . . 23

2 Illumination invariance **25**

3 Pose-invariant body recognition **27**

4 Development of domain-specific regions **28**

5 Methods **32**
5.1 Stimuli . . . 32
5.2 The test of transformation-tolerance from a single example view . . . 32
5.3 Measuring transformation compatibility (ψ) . . . 33
5.4 Clustering by transformation compatibility . . . 33
5.5 Evaluating the clustered models on subordinate-level and basic-level tasks . . . 34

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 A theory of architectures for invariant object recognition

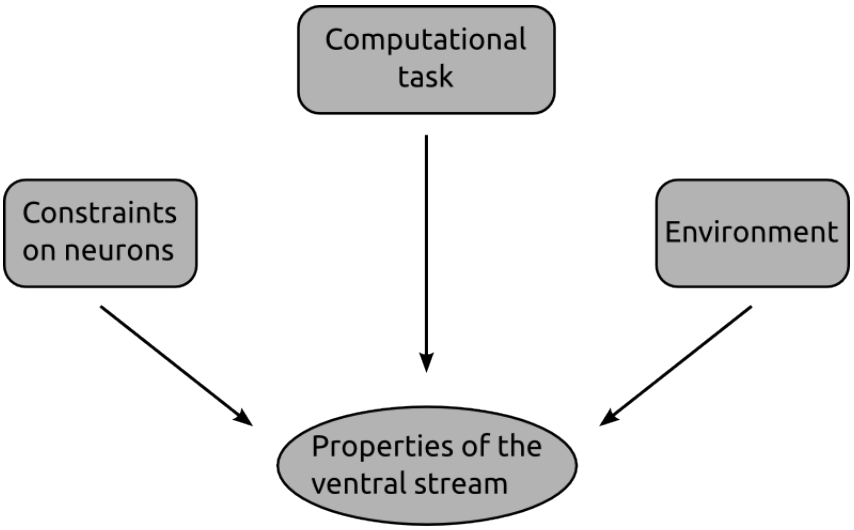

Figure 7: It is hypothesized that properties of the ventral stream are determined by these three factors. We are not the only ones to identify them in this way. For example, Simoncelli and Olshausen distinguished the same three factors [67]. The crucial difference between their *efficient coding hypothesis* and our *invariance hypothesis* is the particular computational task that we consider. In their case, the task is to provide an efficient representation of the visual world. In our case, the task is to provide an invariant signature supporting object recognition.

The new theory of architectures for object recognition [1]—applied here to the ventral stream—is quite general. It encompasses many non-biological hierarchical networks in the computer vision literature in addition to ventral stream models like HMAX. It also implies the existence of a wider class of hierarchical recognition algorithms that has not yet been fully explored. The conjecture with which this paper is concerned is that the algorithm implemented by the ventral stream's feedforward processing is in this class. The theory can be developed from four postulates: 1. Computing a representation that is unique to each object and invariant to identity-preserving transformations is the main computational problem to be solved by an object recognition system—i.e., by the ventral stream. 2. The ventral stream's feedforward, hierarchical operating mode is sufficient for recognition [78, 26, 29]. 3. Neurons can compute high-dimensional dot products between their inputs and a stored vector of synaptic weights [50]. 4. Each layer of the hierarchy implements the same basic "HW-"module, performing filtering and pooling operations via the scheme proposed by Hubel and Wiesel for the wiring of V1 simple cells to complex cells [25].

We argue that as long as these postulates are approximately correct, then the algorithm implemented by the (feedforward) ventral stream is in the class described by the theory, and this is sufficient to explain its domain-specific organization.

1.1 The first regime: generic invariance

First, consider the (compact) group of 2D in-plane rotations G. With some abuse of notation, we use g to indicate both an element of G and its unitary representation acting on images. The orbit of an image I under the action of the group is $O_I = \{gI | g \in G\}$. The orbit is invariant and unique to the object depicted in I. That is, $O_I = O_{I'}$ if and only if $I' = gI$ for some $g \in G$. For an example, let I be an image. Its orbit O_I is the set of all images obtained by rotating I in plane. Now consider, $g_{90^\circ} I$, its rotation by 90°. The two orbits are clearly

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the same $O_I = O_{g_{90^\circ} I}$—the set of images obtained by rotating I is the same as the set of images obtained by rotating $g_{90^0} I$.

The fact that orbits are invariant and unique (for compact groups) suggests a recognition strategy. Simply store the orbit for each object. Then when new objects appear, check what orbit they are in. While that strategy would certainly work, it would be impossible to implement in practice. If we were restricted to cases where we had already stored the entire orbit then we would only ever be able to recognize objects that we had previously encountered under all their possible appearances. The key property that enables this approach to object recognition is the following condition. For a stored *template* t

$$\langle gI, t\rangle = \langle I, g't\rangle \quad \exists g' \in G \;\; \forall g \in G. \tag{4}$$

It is true whenever g is unitary since in this case $g' = g^{-1}$. It implies that it is not necessary to have the orbit of I before the test. Instead, the orbit of t is sufficient. Eq. (4) enables the invariance learned from observing a set of templates to transfer to new images. Consider the case where the full orbit of several templates $t^1, \ldots, t^K$ were stored, but I is a completely novel image. Then an invariant signature $\mu(\cdot)$ can be defined as

$$\mu(I) = \begin{pmatrix} P_g\left(\left\langle I, gt^1\right\rangle\right) \\ \vdots \\ P_g\left(\left\langle I, gt^K\right\rangle\right) \end{pmatrix} \tag{5}$$

Just as in the HMAX case, P_g must be unchanged by permuting the order of its arguments, e.g., $P_g(\cdot) =$ max$_g(\cdot)$ or $\sum_g(\cdot)$.

So far, this analysis has only applied to compact groups. Essentially the only interesting one is in-plane rotation. We need an additional idea in order to consider more general groups—it will also be needed later when we consider non-group transformations in the theory's second regime. The idea is as follows. Most transformations are generally only observed through a range of transformation parameters. For example, in principle, one could translate arbitrary distances. But in practice, all translations are contained within some finite window. That is, rather than considering the full orbit under the action of G, we consider partial orbits under the action of a subset $G_0 \subset G$ (note: G_0 is not a subgroup). We can now define the basic module that will repeat through the hierarchy. An HW-module consists of three elements: (t, G_0, μ). The "response" of an HW-module is $\mu(I) = P_{g\in G_0}(\langle I, gt\rangle)$. Note that if G_0 is a set of translations and $P_g(\cdot) =$ max$_g(\cdot)$, then one such HW-module is exactly equivalent to an HMAX C-unit (defined in the main text). The subset G_0 can be thought of as the pooling domain, for the case of translation, it has the same interpretation as a spatial region as in HMAX.

Consider, for simplicity, the case of 1D images (centered in zero) transforming under the 1D locally compact group of translations. What are the conditions under which an HW-module will be invariant over the range $G_0 = [-b, b]$? Let $P_g(\cdot) := \sum_{x\in[-b,b]} \eta(\cdot)$, where η is a positive, bijective function. The signature vector components will be given by

$$\mu^k(I) = \sum_{x\in[-b,b]} \eta\Big(\left\langle I, T_x t^k\right\rangle\Big)$$

where T_x is the operator acting on a function f as $T_x f(x') = f(x' - x)$. Suppose we transform the image I (or equivalently, the template) by a translation of $\bar{x} > 0$, implemented by $T_{\bar{x}}$. Under what conditions does $\mu^k(I) = \mu^k(T_{\bar{x}} I)$? Note first that $\left\langle I, T_x t^k\right\rangle = (I * t^k)(x)$, where $*$ indicates convolution. By the properties of the convolution operator, we have $[(T_{\bar{x}} I) * t^k](x) = T_{\bar{x}}(I * t^k)(x)$ which implies

$$\text{supp}[(T_{\bar{x}} I) * t^k] = T_{\bar{x}}\text{supp}(I * t^k).$$

This observation allows us to write a condition for the invariance of the signature vector components with respect to the translation $T_{\bar{x}}$ (see also Fig. 8). For a positive nonlinearity η, (no cancelations in the sum) and bijective (the support of the dot product is unchanged by applying η) the condition for invariance is:

$$T_{\bar{x}}\text{supp}(\left\langle I, T_x t^k\right\rangle) \subseteq [-b, b] \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

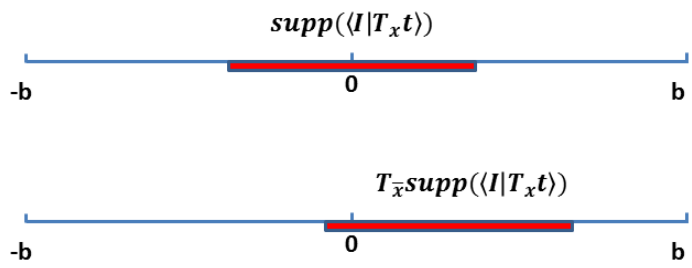

Figure 8: Localization condition of the S-unit response for invariance under the transformation $T_{\bar{x}}$

Eq. 6 is a localization condition on the S-unit response. It is necessary and sufficient for invariance. In this case, eq. (4) is trivial since we are considering group transformations.

Remark: The exposition of the theory given here is only sufficient for understanding a particular special case. In the general case [1], we allow each "element" of the signature (as defined here) to be a vector representing a distribution of one-dimensional projections of the orbit.

1.2 The second regime: class-specific invariance

So far, we have explained how the localization properties of the S-response allow invariance in the case of partially observed group transformations. Next, we show how localization still enables approximate invariance (ϵ-invariance) even in the case of non-group (smooth) transformations. However, as will be shown below, in order for eq. (4) to be (approximately) satisfied, the class of templates needs to be much more constrained than in the group case.

Consider a smooth transformation parametrized by $r \in \mathbb{R}$, T_r; the Taylor expansion of $T_r I$ w.r.t. r around, e.g., zero is:

$$T_r(I) = T_0(I) + J^I(I)r + O(r^2) = I + J^I(I)r + O(r^2) = L_r^I(I) + O(r^2). \tag{7}$$

where J^I is the Jacobian of the transformation T, and $L^I(\cdot) = e(\cdot) + J^I(\cdot)r$. The operator L^I corresponds to the best linearization around the point $r = 0$ of the transformation T_r. Let R be the range of the parameter r such that $T_r(I) \approx L_r^I(I)$. If the localization condition holds for a subset of the transformation parameters contained in R, i.e.

$$\left\langle T_r I, t^k \right\rangle \approx \left\langle L_r^I I, t^k \right\rangle = 0, \quad r \notin R \tag{8}$$

and as long as the pooling range P, in the r parameter is chosen so that $P \subseteq R$, then we are back in the group case, and the same reasoning used above for translation still holds.

However this is not the case for eq. (4). The tangent space of the image orbit is given by the Jacobian, and it clearly depends on the image itself. Since the tangent space of the image and of the template will generally be different (see fig. 9), this prevents eq. (4) from being satisfied. More formally, for $r \in R$:

$$\left\langle L_r^I(I), t^k \right\rangle = \left\langle I, [L_r^I]^{-1} t^k \right\rangle \quad \Leftrightarrow \quad L_r^I = L_r^{t^k}.$$

That is, eq. (4) is only satisfied when the image and template "transform the same way" (see Fig. 9).

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

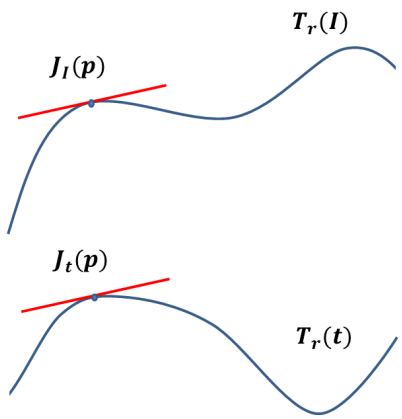

Figure 9: The Jacobians of the orbits of the image around the point p and the template must be approximately equal for eq. (4) to hold in the case of smooth transformations.

To summarize, the following three conditions are needed to have invariance for non-group transformations:

1. The transformation must be differentiable (the Jacobian must exist).
2. A localization condition of the form in eq. (8) must hold to allow a linearization of the transformation.
3. The image and templates must transform "in the same way", i.e. the tangent space of their orbits (in the localization range) must be equal. This is equivalent to $J^I \equiv J^{t^k}$.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Illumination invariance

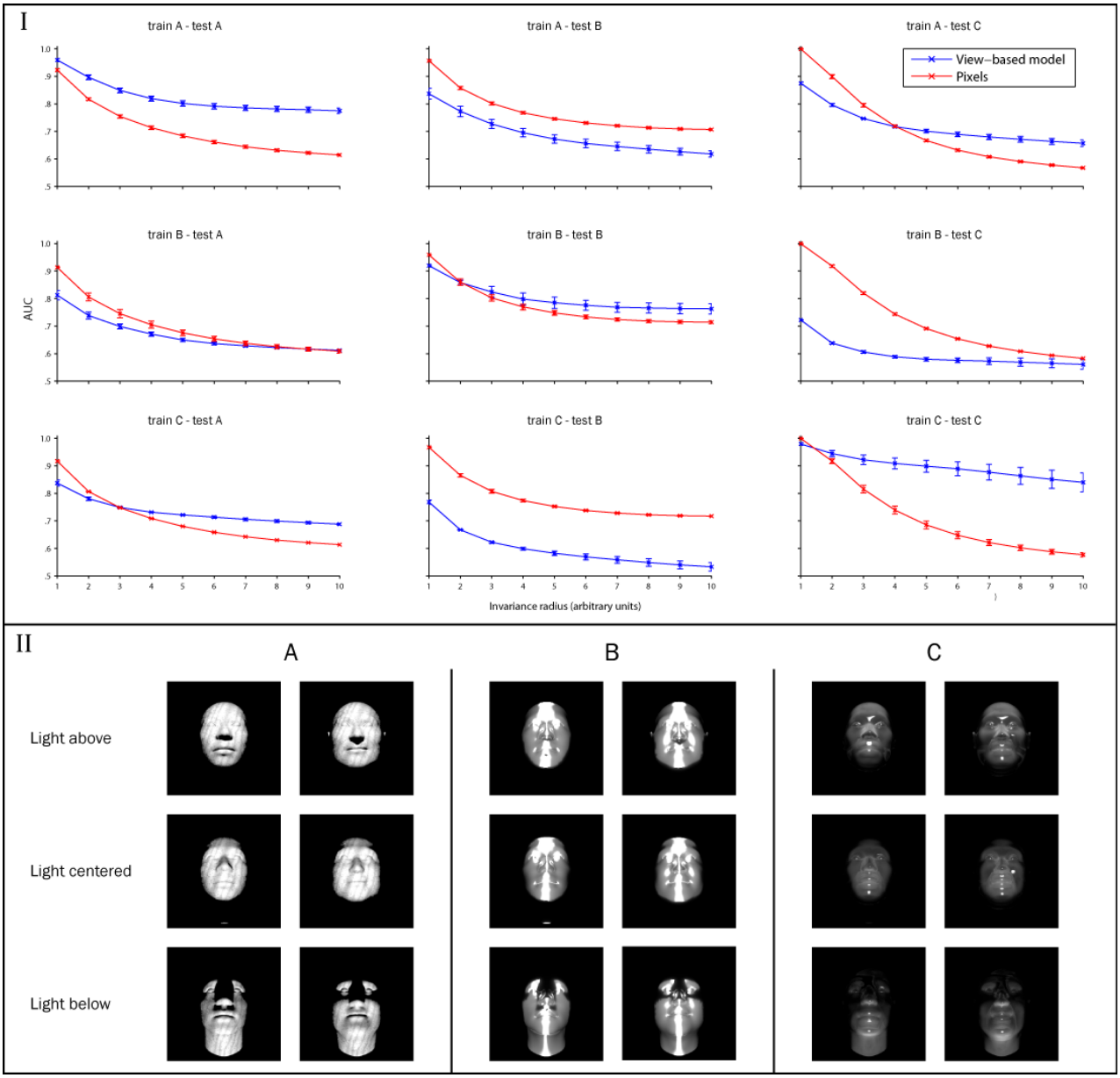

Figure 10: Class-specific transfer of illumination invariance. Bottom panel (II): Example images from the three classes. Top panel (I): The left column shows the results of a test of illumination invariance on statues of heads made from different materials (class A), the middle column shows results for class B and the right column shows the results for class C. The view-based model (blue curve) was built using images from class A in the top row, class B in the middle row, and class C in the bottom row. The abscissa of each plot shows the maximum invariance range (arbitrary units of the light source's vertical distance from its central position) over which target and distractor images were generated. The view-based model was never tested on any of the images that were used as templates. Error bars (+/- one standard deviation) were computed over 20 cross validation runs using different choices of template and test images.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Illumination is also a class-specific transformation. The appearance of an object after a change in lighting direction depends both on the object's 3D structure and on its material properties (e.g. reflectance, opacity, specularities). Figure 10 displays the results from a test of illumination-invariant recognition on three different object classes which can be thought of as statues of heads made from different materials—A: wood, B: silver, and C: glass. The results of this illumination-invariance test follow the same pattern as the 3D rotation-invariance test. In both cases the view-based model improves the pixel-based models' performance when the template and test images are from the same class (fig. 10—plots on the diagonal). Using templates of a different class than the test class actually lowered performance below the pixel-based model in some of the tests e.g. train A–test B and train B–test C (fig. 10—off diagonal plots). This simulation suggests that these object classes have high $\overline{\psi}$ respect to illumination transformations. However, the weak performance of the view-based model on the silver objects indicates that it is not as high as the others (see the table below). This is because the small differences in 3D structure that define individual heads give rise to more extreme changes in specular highlights under the the transformation.

Object class	Transformation	$\overline{\psi}$
Glass statues	illumination	0.56320
Sliver statues	illumination	0.35530
Wood statues	illumination	0.53990

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Pose-invariant body recognition

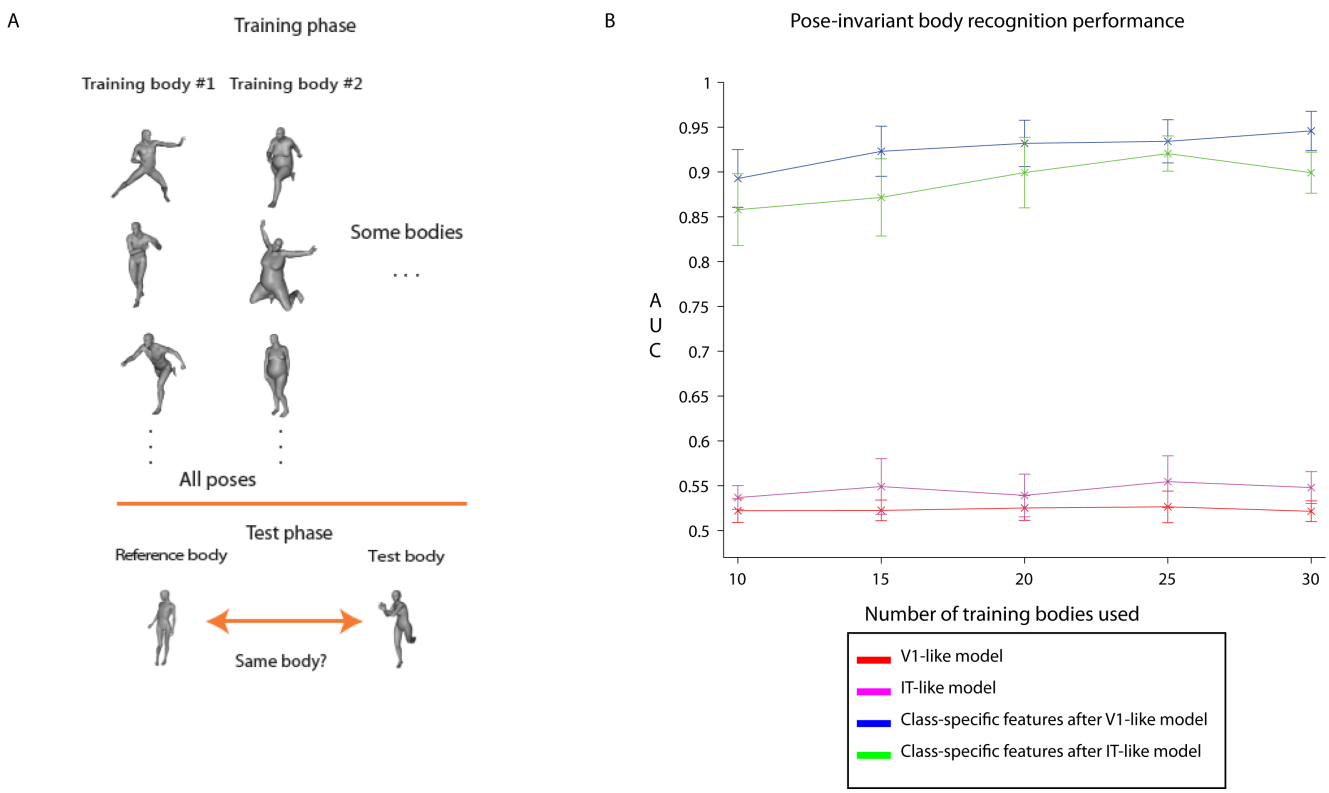

Figure 11: **A.** Example images for the pose-invariant body-recognition task. The images appearing in the training phase were used as templates. The test measures the model's performance on a same-different task in which a reference image is compared to a query image. 'Same' responses are marked correct when the reference and query image depict the same body (invariantly to pose-variation).
B. Model performance: area under the ROC curve (AUC) for the same-different task with 10 testing images. The X-axis indicates the number of bodies used to train the model. Performance was averaged over 10 cross-validation splits. The error bars indicate one standard deviation over splits.

Let $B = \{b_1, b_2, ..., b_n\}$ be a set of bodies and $P = \{p_1, p_2, ..., p_n\}$ be a set of poses. Let d be the dimensionality of the images. We define the rendering function $t_p : B \to \mathbb{R}^d$. In words, we say $t_p[b]$ renders an image of body b in pose p. In that case the argument b is the template and the subscript p indicates the transformation to be applied.

We obtain the signature vector $\mu : X \to \mathbb{R}^m$ by pooling the inner products of the input image with different renderings of the same template.

$$\mu(x) = \begin{pmatrix} \max(\langle I, t_1(\tau_1)\rangle, \langle I, t_2(\tau_1)\rangle, \ldots, \langle I, t_n(\tau_1)\rangle) \\ \max(\langle I, t_1(\tau_2)\rangle, \langle I, t_2(\tau_2)\rangle, \ldots, \langle I, t_n(\tau_2)\rangle) \\ \vdots \\ \max(\langle I, t_1(\tau_m)\rangle, \langle I, t_2(\tau_m)\rangle, \ldots, \langle I, t_n(\tau_m)\rangle) \end{pmatrix} \tag{9}$$

As in some HMAX implementations (e.g., Serre et al. (2007) [66]), we used a Gaussian radial basis

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

function for the S-unit response. It has similar properties to the normalized dot product.

$$\langle I, t_i(\tau_j)\rangle = \exp\{\sigma * \sum((I - t_i(\tau_j))^2)\} \tag{10}$$

Where σ is the Gaussian's variance parameter.

The class-specific layer takes in any vector representation of an image as input. We investigated two hierarchical architectures built off of different layers of the HMAX model (C1 and C2-global) [66]—referred to in fig. 11 as the V1-like and IT-like models respectively.

For the pose-invariant body recognition task, the template images were drawn from a subset of the 44 bodies—rendered in all poses. In each of 10 cross-validation splits, the testing set contained images of 10 bodies that never appeared in the model-building phase—again, rendered in all poses (fig. 11).

The HMAX models perform almost at chance. The addition of the class-specific mechanism significantly improves performance on this difficult task. That is, models without class-specific features were unable to perform the task while class-specific features enabled good performance on this difficult invariant recognition task (fig. 11).

Downing and Peelen (2011) argued that the extrastriate body area (EBA) and fusiform body area (FBA) "jointly create a detailed but cognitively unelaborated visual representation of the appearance of the human body". These are perceptual regions—they represent body shape and posture but do not explicitly represent high-level information about "identities, actions, or emotional states" (as had been claimed by others in the literature [14]). The model of body-specific processing suggested by the simulations presented here is broadly in agreement with this view of EBA and FBA's function. It computes, from an image, a body-specific representation that could underlie many further computations e.g. action recognition, emotion recognition, etc.

4 Development of domain-specific regions

The goal of this section is to illustrate in detail the logic through which the invariance hypothesis could be used to make predictions for the specific object classes that will "get their own private piece of real estate in the brain" [32]. As such, we begin by enumerating the extra assumptions we use in this section, beyond the standard ones of the theory (see section 1).

1. For simplicity, we consider only out-of-plane rotations here. Quite similar simulations could be done for other class-specific transformations.

2. Over the course of development (or evolution), the orbits of objects are stored in cortex (as HW-modules). The distribution of objects/transformations encountered under natural visual experience determines which HW-modules are stored, but it has no influence on the specific cortical location of their storage.

3. The arrangement of HW-modules on cortex is related to an intrinsic property of the orbits they encode. Here we assume a relationship between $\psi(A, B)$ and cortical proximity (see main text discussion for justification).

Assumptions 1-3 suffice to predict clusters of HW-modules for different templates. We need one more additional assumption to come to a prediction about which regions should appear in fMRI experiments.

4. The number of HW-modules in a cluster and the proportion belonging to different categories determine the predicted BOLD response for contrasts between the categories. For example, a cluster with 90% face HW-modules, 10% car HW-modules, and no other HW-modules would respond strongly in the faces - cars contrast, but not as strongly as it would in a faces - airplanes contrast. We assume that clusters containing very few HW-modules are too small to be imaged with the resolution of fMRI—though they may be visible with other methods that have higher resolution.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

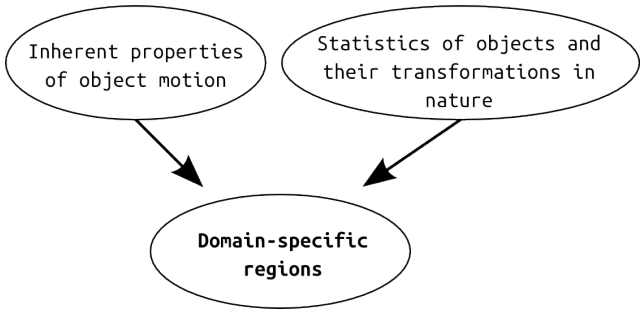

Figure 12: Two factors are conjectured to influence the development of domain-specific regions.

For now, we are mostly interested in the influence of transformation compatibility, in part because it is the novel part of our proposal, also because we don't have good estimates of the distribution of objects (and their transformations) encountered under natural vision. Though a serious study of those statistics is now motivated, such a project is beyond the scope of the present paper.

We consider three different arbitrary choices for the distributions of objects from five different categories: faces, bodies, vehicles, chairs, and animals (see table 1). Importantly, one set of simulations used statistics which were strongly biased against the appearance of faces as opposed to other objects.

	Name of simulation	Faces	Bodies	Animals	Chairs	Vehicles
A.	"Realistic"	76	32	16	16	16
B.	Uniform	30	30	30	30	30
C.	Biased against faces	16	32	36	36	36

Table 1: Numbers of objects used for each simulation. In the "realistic" simulation, there were proportionally more faces.

By assumption #3, for any two learned objects A and B, the distance $d(A, B)$ along cortex between their respective HW-modules is proportional to their pairwise transformation compatibility, i.e.

$$d(A, B) \propto \psi(A, B) \tag{11}$$

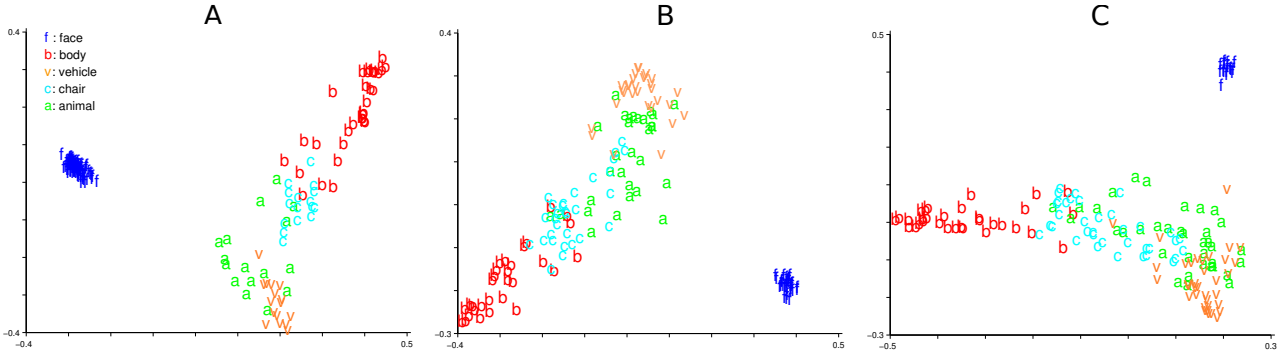

Figure 13: Multidimensional Scaling (MDS) [79] visualizations of the object sets under the $\psi(A, B)$-dissimilarity metric for the three object distributions: A. "realistic", B. uniform, and C. biased against faces (see table 1).

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The clustering algorithm, detailed in section 5.4, can be summarized as follows: Consider an object recognition system with a number of bins corresponding to sets of HW-modules. When an object is learned, add its newly-created HW-module to the bin with which its transformations are most compatible. If the new object's average compatibility with all the existing bins is below a certain threshold, then create a new bin to hold the HW-module of the newly learned object. To model visual development, this procedure is repeated many times with new objects—sampled according to the distribution of objects encountered in natural vision (or whatever approximation is available). At the end of the development process, we can identify each bin with a domain-specific region.

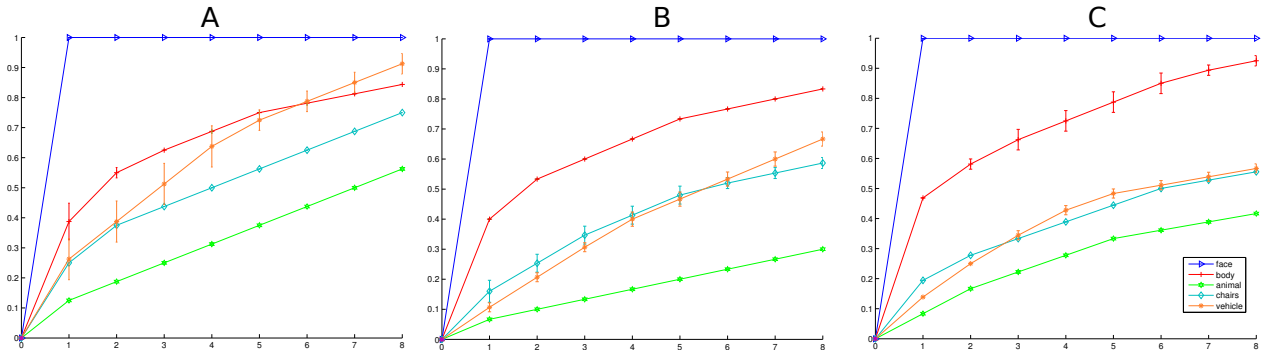

Figure 14: The percentage of objects in the first N clusters containing the dominant category object (clusters sorted by number of objects in dominant category). A, B and C are respectively, the "realistic" distribution, uniform distribution, and the biased against faces distribution (see table 1)). 100% of the faces go to the first face cluster—only a single face cluster developed in each experiment. Bodies were more "concentrated" in a small number of clusters, while the other objects were all scattered in many clusters—thus their curves rise slowly. These results were averaged over 5 repetitions of each clustering simulation using different randomly chosen objects.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

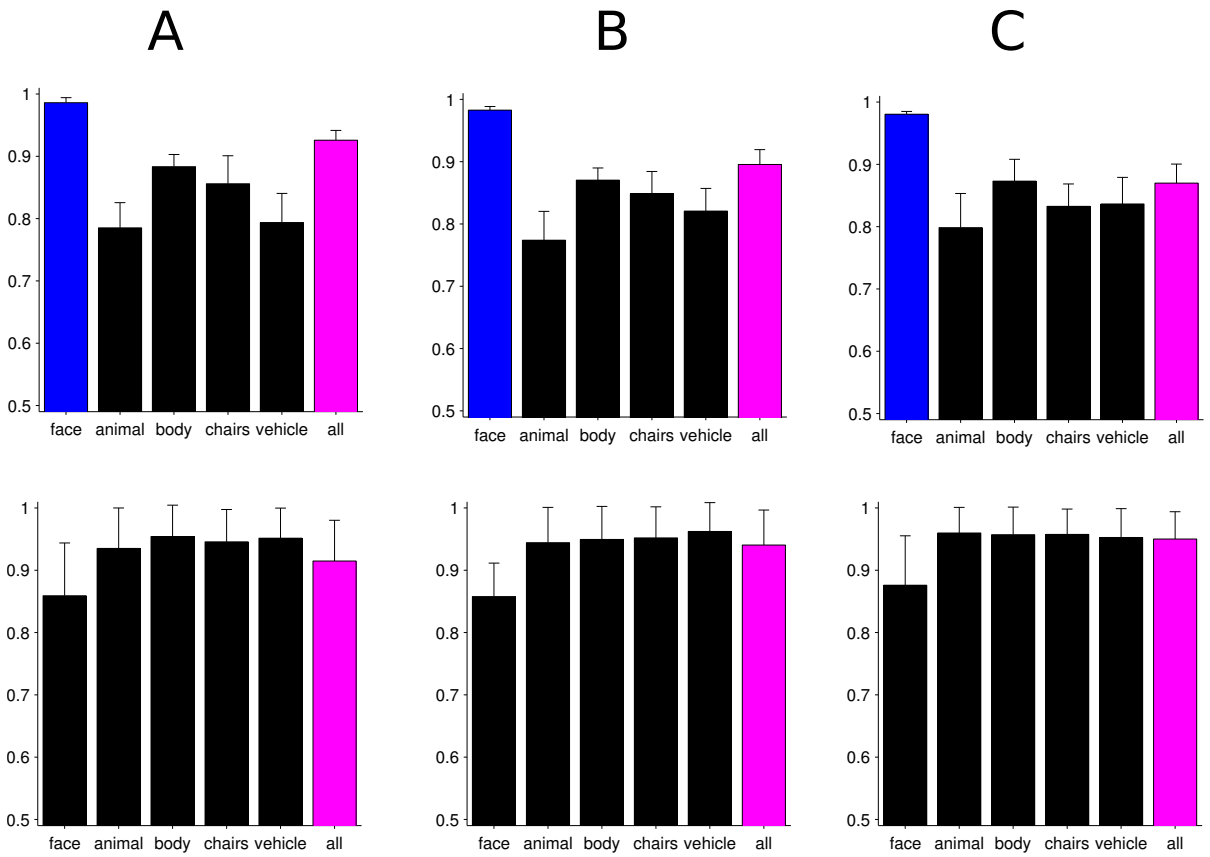

Figure 15: The classification performance on face recognition, a subordinate-level task (top row) and car vs. airplane, a basic-level categorization task (bottom row) using templates from each cluster. 5-fold cross-validation, for each fold, the result from the best-performing cluster of each category is reported. A, B and C indicate "realistic", uniform, and biased distributions respectively (see table 1). Note that performance on the face recognition task is strongest when using the face cluster while the performance on the basic-level car vs. airplane task is not stronger with the vehicle cluster (mostly cars and airplanes) than the others.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5 Methods

5.1 Stimuli

Faces and novel object classes

All objects were rendered with perspective projection. For translation experiments: there were 100 faces and 100 random noise patterns randomly partitioned into sets of 30 for templates and 30 for testing. For rotation in depth experiments: there were 40 untextured faces, 20 class B, and 20 class C objects, randomly picked 10 template objects and 10 test objects for each run of the experiment, repeated 20 times with different randomly chosen objects. Each face/object was rendered (using Blender [6]) at each orientation in 5° increments from -95° to 95°. The untextured face models were generated using Facegen [68].

Illumination

Illumination: Within each class the texture and material properties were exactly the same for all objects. We used Blender to render images of each object with the scene's sole light source placed in different locations. The 0 position was set to be in front of the object's midpoint; the light was translated vertically. The most extreme translations brought the light source slightly above or below the object. We obtained the material data files from the Blender Open Material Repository (http://matrep.parastudios.de/). We rendered images of 40 heads with each material type and and randomly picked 20 to be templates and 20 for testing in each of 20 cross validation runs.

Bodies / pose

DAZ 3D Studio was used to render each of 44 different human bodies under 32 different poses, i.e., 44*32 =1408 images in total.

Objects for the development (iterative clustering) experiments

Blender was used to render images of 3D models from the Digimation archive (platinum edition) from a range of viewpoints: -90° to 90° in increments of 5 degrees. A set of textured face models generated with FaceGen were added to the Digimation set.

5.2 The test of transformation-tolerance from a single example view

We simulated tests of initial invariance for unfamiliar faces. The specific task we modeled is a same-different task. In human behavior, it would correspond to a task where the subject is first shown a reference image and then asked to compare it to a query image. The query may be an image of the same face as the reference (the target), or it may depict a distractor face. In either case, the query image may be transformed. For example, in one trial, the task could be to recognize "Tommy's face"—oriented 0° in the reference image—versus distractor images of other people's faces. Both target and distractor query images might be rotated away from the reference view.

This task is modeled using a nearest-neighbors classifier. The reference image's signature is chosen to be the center. The classifier then ranks all the query images' signatures by their distance from the reference. We vary the threshold for which the classifier will respond 'same' to compute a bias-free measure of performance (AUC)—analogous to d' for the corresponding behavioral experiment [41, 21]. Figures 2 and 3 shows the AUC computed for a range of task difficulties (the abscissa). These figures show how discriminability declines as the range of transformations applied to the query images is widened. A task at a larger invariance range subsumes all the tasks at smaller invariance ranges; thus the discriminability curves can never increase as the invariance range is widened. A flat AUC curve indicates that discriminability is unaffected by the transformation. That is, it indicates that the model is invariant to that transformation.

For the body-pose invariance experiments, the task remained the same, but the way of reporting results changed since the transformation was not parameterized.

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.3 Measuring transformation compatibility (ψ)

Let A_i be the i_{th} frame of the video of object A transforming and B_i be the i_{th} frame of the video of object B transforming. We define a compatibility function $\psi(A, B)$ to quantify how similarly objects A and B transform.

First, approximate the Jacobian of a transformation sequence by the "video" of difference images: $J_A(i) = A_i - A_{i+1}$ $(\forall i)$.

Then we can define the transformation compatibility as:

$$\psi(A, B) = \text{Mean}_i(\text{similarity}(J_A(i), J_B(i))) \tag{12}$$

Similarity was measured using a normalized dot product.

Transformation compatibility can be visualized by Multidimensional Scaling (MDS) [79]. The input to the MDS algorithm is the pairwise "similarity matrix" containing the transformation compatibilities between all pairs of objects.

5.4 Clustering by transformation compatibility

The pseudocode for our iterative clustering algorithm is given below (algorithm 1). We define the transformation compatibility $\overline{\psi}$ of a cluster to be the average of the pairwise compatibilities $\psi(A, B)$ of all objects in the cluster

$$\overline{\psi} := \text{mean}(\psi(A, B)) \;\; \text{for all pairs of objects } (A, B) \text{ from a cluster.} \tag{13}$$

bioRxiv preprint doi: https://doi.org/10.1101/004473; this version posted April 24, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Algorithm 1 Iterative clustering to model ventral stream development

```
Input: All Objects: O, i_th Object: O_i where i = 1...N, Threshold: T)
Output: ClusterLabels
Code:
ClusterLabels(1) = 1
ψ̄ = computeCompatibility(ClusterLabels)
for i = 2 to N do
  ψ = computeCompatibilityWithEveryCluster(i, O, ClusterLabels)
  [MaxValue MaxIndex] = max(ψ)
  if MaxValue > T then
    ClusterLabels(i) = MaxIndex //Assign to the cluster with the highest compatibility.
  else
    ClusterLabels(i) = max(ClusterLabels) + 1 //Create a new cluster
  end if
  ψ̄ = updateCompatibility(ψ, CurrentClusterCompatibility, ClusterLabels(i))
end for

Function computeCompatibilityWithEveryCluster(IDX,AllObjects,ClusterLabels)
//Initialize ψ as an empty array of length #Clusters.
for i = 1 to #Clusters do
  Objects = GetObjectsFromCluster(i, AllObjects, ClusterLabels)
  for j = 1 to #Objects do
    tmpArray(j) = compatibilityFunction(AllObjects(IDX), Objects(j))
  end for
  ψ(i) = mean(tmpArray);
end for
Return ψ
EndFunction
```

5.5 Evaluating the clustered models on subordinate-level and basic-level tasks

We evaluated the performance of the models trained with templates from each cluster on a subordinate face verification task (same-different matching) and a basic-level car vs. airplane verification task. In the face verification task, we ran 5-fold validation, each fold contains 48 training and 12 testing faces. For the basic level categorization task , we ran 5-fold validation with 96 (48+48) training and 24 (12+12) testing objects in each fold. For both tasks, 4000 training and 4000 (independent) testing pairs were used for training and testing the classifier.