The Invariance Hypothesis and the Ventral Stream

by

Joel Zaidspiner Leibo

B.S./M.S. Brandeis University 2008

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2014

Author . . . . . .
Department of Brain and Cognitive Sciences
September 5, 2013

Certified by . . . . . .
Tomaso Poggio
Eugene McDermott Professor
Thesis Supervisor

Accepted by . . . . . .
Matthew Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

The Invariance Hypothesis and the Ventral Stream

by

Joel Zaidspiner Leibo

Submitted to the Department of Brain and Cognitive Sciences
on September 5, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The invariance hypothesis is the claim that properties of the ventral stream can be deduced from 1. a consideration of the computational problem with which it is engaged: invariant object recognition, 2. constraints of the neural "hardare", and 3. the visual environment. We are interested here in a particular instantiation of this idea. A recent general theory of hierarchical networks for invariant recognition [1] describes many modern convolutional networks as special cases, and also implies the existence of a wider class of algorithms, which we are only now beginning to explore. Our version of the invariance hypothesis is the claim that the algorithm implemented by the ventral stream is also in this class. As applied to the brain, the theory follows from a few simple and commonly accepted premises. This thesis contributes several models/studies in which properties of the ventral stream are deduced and explained in the context of the theory. The main contribution here is providing a general framework through which disparate results concerning many parts of the ventral stream, and even different levels of analysis [2], can be bridged and understood. In that sense, it is primarily a Neuroscience contribution. However, the ideas and algorithms it suggests may also have implications for the broader question of how to learn representations capable of supporting intelligence.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor

Acknowledgments

Paul Krugman's Ph.D. thesis begins with the words:

"The ideas in this thesis are largely the product of discussions with faculty and fellow students, so that I am no longer sure which ideas, if any, can properly be considered mine".

The same is true here.

This thesis emerged from conversations between myself and the other members of the "magic" group: Tomaso Poggio, Fabio Anselmi, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and our most recent member: Qianli Liao.

First, particular thanks goes to my advisor Tomaso Poggio—despite my frequent interruption of his speaking—I always learn something valuable from our interaction. I hope to have understood his intellectual style well-enough to pass along a not-too-badly-deformed version to my own students. I have certainly been trying! One of my favorite parts of graduate school has been mentoring undergraduate and M.Eng. students: Sara Brockmueller, Mark Rogers, Heejung Kim, Darrel R. Deo, Jeremy Wohlwend, and Qianli Liao. As their designated supervisor, I was pleased to find them excelling in unsupervised learning—and only rarely in their training has error backpropagation been necessary.

Thanks to the rest of my thesis committee: Nancy Kanwisher, Pawan Sinha, and Winrich Freiwald. My annual committee meetings have been useful exercises, and not too stressful. I appreciated all your insights and feel influenced by your work. I'd like to thank Professor Freiwald in particular for all his, quite significant, contributions to writing and idea-refinement for the mirror-symmetry project (chapter 5).

I was also fortunate to have four other influential mentors who were not on my thesis committee: James DiCarlo, Josh Tenenbaum, Lorenzo Rosasco, and Demis Hassabis. Conversations with all four of them have been highly influential in shaping my—ever evolving—thoughts on which questions and research directions are really fundamental.

Thanks also to my other collaborators, particularly Leyla Isik and Ethan Meyers.

Though we don't always agree, the debate is always interesting, and in Ethan's case it's valuable to see the case made for an extreme empiricist viewpoint, a contrast with my usually more theoretical outlook.

Thanks to Gadi Geiger and Steve Voinea for always making sure there was a non-dairy meal for me at group meetings. It was much appreciated!

Thanks to all my friends at MIT and elsewhere! Especially, those that helped me out during various periods of homelessness (surprising how many of these I had in graduate school): Jonathan Lansey, Maxine Hesch, Emily Harrison, Beth Zweig, Rocky Acosta, Jiyeon Woo and Kaan Erbay. Also Fabio Anselmi, he doesn't know it, but I sometimes sleep on the bean bags in the room adjoining his office. I also want to thank several other recently close friends: Nuné and Martin Lemaire, Kelly Geyer, Sangyu Xu, Josh Manning, Courtney DeMaria, Matt Greene, Maria Berezina, Molly Sicchio, and all my friends from the Wednesday CBC meetups. I'd also like to thank Brant Greishaber for attempting to teach me to play the guitar, Tony Watt for attempting to teach me to play the mandolin, and Lydia Zotto for attempting to explain how a harp works.

Some of the most important people I've interacted with in graduate school have been Kathleen Sullivan and Denise Heintze. Every graduating student from CBCL and BCS likely owes more to them than we even know about. Thanks for everything!

Finally, I'd like to thank my family. None of this would be possible without their support.

Contents

Chapter 1

Introduction

Unlike Athena, the new ventral stream theory foreshadowed in these pages did not spring fully-formed from the head of Zeus. We: primarily Tomaso Poggio, Fabio Anselmi, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and myself, developed it—and continue to refine it—within a context in which the questions considered in this dissertation loom large. Each of its four main chapters can be read independently, but a common thread runs through them. In other manuscripts—some already released, and some currently in preparation—we spin the same thread into the cloth of the new theory. Most of the chapters of this dissertation propose (or analyze) specific models designed to elucidate particular aspects of the ventral stream and the object recognition algorithm it implements. Some of them (chapters two and three) were mostly completed before we knew they were parts of a larger story. Whereas, the work of chapters four and five was undertaken more consciously as part of the larger theory's development. A version of each chapter has appeared, or soon will appear, as a standalone article.

Chapter one, the introduction, has two parts. The first part gives general background on learning and perception. The second part introduces the theory of the ventral stream which is used throughout this thesis—particularly in its last two chapters.

Chapter two, Measuring invariance, proposes an operational definition of invariance which can be used to compare neural data, behavior, and computational

models. It then describes the application of the definition to an analysis of the invariance properties of HMAX, an older model of the ventral stream [3, 4], which motivated much of our later work. In particular, we showed that, from a certain point of view, there really is no selectivity-invariance trade-off. HMAX can be almost perfectly invariant to translation and scaling. Furthermore, we showed that high similarity of the template images to the to-be-recognized images is not a requirement for invariance and that, in many cases, surprisingly small numbers of templates suffice for robust recognition despite translation and scaling transformations. The finding that random dot pattern templates are perfectly reasonable templates to use for invariant recognition was surprising at the time, but thanks to subsequent development of the theory, is now well-understood. Other groups made similar observations around the same time [5, 6]. A version of chapter two appeared as a CSAIL technical report in 2010: [7], and we also published some of the same ideas in [8].

Chapter three, Learning and disrupting invariance in visual recognition with a temporal association rule, was joint work. Leyla Isik and I contributed equally. We modeled Li and DiCarlo's "invariance disruption" experiments [9, 10] using a modified HMAX model. In those experiments, monkeys passively viewed objects which changed identity while saccades brought them from a peripheral retinal position to the fovea. As little as one hour of exposure to this strange visual environment caused significant effects on the position invariance of single units. But don't "errors of temporal association" like this happen all the time over the course of normal vision? Lights turn on and off, objects are occluded, you blink your eyes—all of these should cause errors in temporal association. If temporal association is really the method by which invariance to larger patterns is developed and maintained, then why doesn't the fact that its assumptions are so often violated lead to huge errors in invariance? In this work, we showed that these models are actually quite robust to this kind of error. As long as the errors are random (uncorrelated), then you can accumulate surprisingly large numbers of them before there is an impact on the performance of the whole system. This

result turns out to be an important one for the later plausibility of our ventral stream theory's reliance on temporal association learning.

Chapter four builds on the results from a paper I wrote for NIPS in 2011 [11] entitled "Why The Brain Separates Face Recognition From Object Recognition". In that paper, we conjectured that the need to discount class-specific transformations (e.g., 3D rotation in depth) is the reason that there are domain-specific subregions (e.g., face patches [12–14]) of the ventral stream. We also give an explanation, based on the new theory, for why the cells in the anterior (high-level) ventral stream are tuned to more complex features than those in the more posterior lower-levels.

Chapter five is concerned with biologically plausible learning rules through which the brain could develop the necessary circuitry to implement these models. As a consequence, we predict that neuronal tuning properties at all levels of the ventral stream are described by the solutions to a particular eigenvalue equation we named the *cortical equation*. In the case of small receptive fields, as in primary visual cortex, its solutions are Gabor wavelets. In the case of large receptive fields, corresponding to a face-specific region, the predicted tuning curves for 3D rotations of faces resemble the mysterious mirror-symmetric tuning curves of neurons in the anterior lateral face patch [15].

1.1 A theory of invariant object recognition

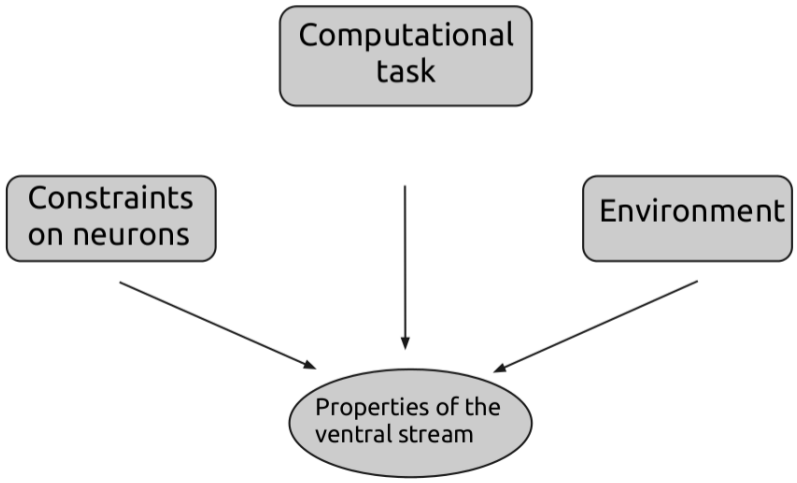

Figure 1-1: Properties of the ventral stream are determined by these three factors. We are not the only ones to identify them in this way. For example, Simoncelli and Olshausen distinguished the same three factors [16]. The crucial difference between their *efficient coding hypothesis* and our *invariance 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 object recognition [1]—applied here to the ventral stream—is very 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 also in this class. The theory is based on four standard 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 [17–19]. 3. Neurons compute high-dimensional dot products between their inputs and a stored vector of synaptic weights [20]. 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 [21]. We do not claim to have identified the ventral stream’s algorithm exactly. Instead, we argue that as long as these postulates are reasonably true, then the algorithm implemented by the (feedforward) ventral stream is in the class described by the theory, and that that is sufficient to explain many of its mysteries.

1.2 High-level perception: recognition and invariance

Visual information passes from the retina along the optic nerve to the thalamus, from there it goes to primary visual cortex (V1): a large region in the extreme posterior part of the brain (occipital lobe). V1 cells are commonly divided into two categories: simple and complex. Hubel and Wiesel, who first made the distinction, found that some V1 cells—which they named simple cells—were optimally tuned to oriented bars. That is, one simple cell may respond when a horizontal line appears in its receptive field while another simple cell may respond when a diagonal line appears in its receptive field. Simple cells are sensitive to the exact location of their preferred stimulus within the receptive field. They have oriented "on" and "off" regions; the appearance of a stimulus in the former increases the cell's firing, while in the latter it suppresses it. Complex cells are also tuned to oriented lines but tolerate shifts in the line's exact position within their receptive field. That is, they have no "off" regions. Most complex cells also have somewhat larger receptive fields than simple cells [21].

Simple cells represent stimuli by an orientation at each position (the coordinates are x,y and angle). Complex cells also represent stimuli by those three coordinates but their spatial sensitivity is diminished. In this sense, the pattern of activity over the population of complex cells can be thought of as a blurred version of the representation carried by the simple cells. V1 inherits the retinotopic organization of its inputs. Thus, cells with receptive fields in nearby regions of the visual field are also located nearby one another in cortex.

Hubel and Wiesel conjectured that complex cells are driven by simple cells [21]. A complex cell tuned to an orientation θ tolerates shifts because it is receives its inputs from a set of simple cells optimally tuned to θ at different (neighboring) positions. In the popular energy model, the response of a complex cell is modeled as the sum of squares of a set of neighboring simple cell responses [22].

In downstream higher visual areas, cells respond to increasingly large regions of space. At the end of this processing hierarchy, in the most anterior parts of the ventral visual system—particularly in the anterior parts of inferotemporal cortex (IT)—there are

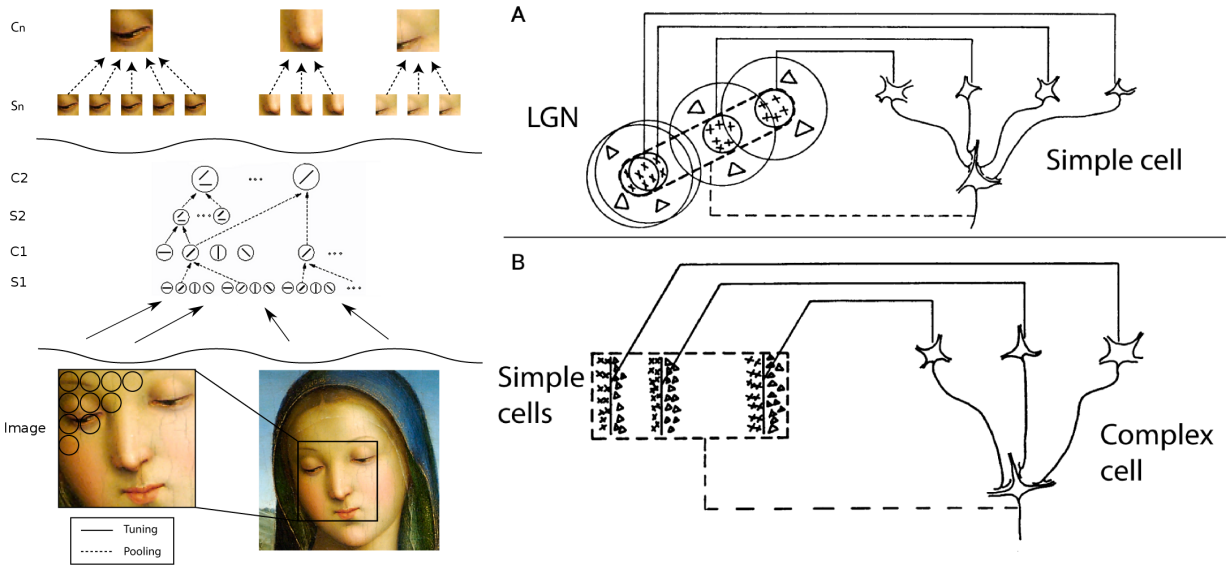

Figure 1-2: *Left: Illustration of a model with a convolutional architecture. In early layers, units represent simple stimuli like oriented bars. In later levels, units are tuned to complex stimuli like patches of natural images. Right: Hubel and Wiesel's illustration of their proposal that complex cells are driven by simple cells. It also shows that simple cell tuning can be produced by a combination of "center-surround" receptive fields like those in the lateral geniculate nucleus (LGN) of the thalamus [21].*

cells that respond invariantly despite significant shifts (up to several degrees of visual angle) [23–25]—but see also [26]. A small object could be translated so that it no longer falls in the receptive fields of any of the photoreceptors that originally imaged it. Even then, the representation in IT could remain largely unchanged. Hubel and Wiesel proposed that a similar organization to the one they discovered in V1 may be repeated beyond early vision; the cells in downstream visual areas could be optimally tuned to more complex image fragments, e.g., corners and T's may come after V1's lines—all the way up to whole object representations in IT.

Fukushima's neocognitron is an early example of a computational model of object recognition built around Hubel and Wiesel's proposal [27]. In that model, alternating layers of "simple" (S) and "complex" (C) cells compute either a "tuning" or a "pooling" operation; other more recent models maintain this organization [3, 4, 28, 29]. S-cells can be thought of as template detectors; they respond when their preferred image (their

template) appears in their receptive field. Since visual areas are approximately retinotopically organized, the pattern of activity of all the cells tuned to a particular template, each at a different location in the visual field—the feature map—can be understood as a map of where in the visual field the template appears. That is, the feature map is the convolution of the inputs with the template. A layer that represents many different features, like V1 with a template for each orientation, is said to contain many feature maps.

C-cells pool together the responses of S-cells at different locations. Thus C-cells are tuned to the same templates as their input S-cells, but they are invariant to stimulus translations within their receptive field. Pooling is typically modeled with a nonlinear operation—as in the energy model: sum of squares of a set of S-cell responses—or the HMAX model: the max of a set of S-cell responses. Since cortex is retinotopically organized, the S feature maps are computed by convolving the template with the inputs. C-feature maps are computed by applying the nonlinear pooling function to local neighborhoods of S cells.

A good question to ask at this point is: how could the proposed organization be wired up through either visual experience or evolution? There are two separate questions here, one is how to learn the templates (the wiring into the S cells), the other is how to learn the pooling (the wiring into the C cells). One answer to the former question is the statistics of commonly viewed image fragments (e.g. [30]). On small spatial scales, low complexity features like the oriented edges of V1 are represented. On larger spatial scales, the cells represent more complex features.

The second problem, that of developing invariance via this scheme, is equivalent to the problem of associating units representing the same template at different positions (and scales) in the visual field. While it is straightforward to achieve this in a computational model, for visual cortex it is one of a class of notoriously difficult correspondence problems. One way the brain might solve it is via temporal-association-based (TAB) methods. The most famous TAB method is Foldiak's trace rule [31]. The trace rule explains how many cells having the same selectivity at different spatial positions could be wired to the same downstream cell by exploiting continuity of motion: cells that fire

to the same stimulus in close temporal contiguity are all presumably selective to the same moving stimulus.

TAB methods are based on the assumption that objects normally move smoothly over time. In the natural world, it is common for an object to appear first on one side of the visual field and then travel to the other side as the organism moves its head or eyes. A learning mechanism that takes advantage of this property would associate temporally contiguous patterns of activity. As a system employing such an algorithm gains experience in the visual world it would gradually acquire invariant template detectors.

Psychophysical studies have tested the temporal association hypothesis by exposing human subjects to altered visual environments in which the usual temporal contiguity of object transformation is violated. Exposure to rotating faces that change identity as they turn around leads to false associations between faces of different individuals [32]. Similarly, exposure to objects that change identity during saccades leads to increased confusion between distinct objects when asked to discriminate at the retinal location where the swap occurred [9].

There is also physiological evidence that the brain uses a TAB method to acquire invariance. Li and DiCarlo showed that exposure to an altered visual environment in which highly-dissimilar objects swap identity across saccades causes anterior IT neurons to change their stimulus preference at the swapped location [10]. They went on to obtain similar results for scale invariance [33]; objects grew or shrank in size and changed identity at a particular scale. After an exposure period, AIT neurons changed their stimulus preference at the manipulated scale. A control unswapped location and scale was unaffected in both experiments.

1.3 Sketch of the general theory of learning invariant representations for object recognition

Note: The primary reference on the theory is the forthcoming paper: Anselmi, Leibo, Mutch, Rosasco, Tacchetti, and Poggio. *Unsupervised Learning of Invariant Representations in Hierarchical Architectures (and in Visual Cortex)* [1]. Until that paper appears the primary reference is [34].

1.3.1 Architecture

It is known that Hubel and Wiesel's principle—that complex cells (C-units) are driven by sets of simple cells (S-units) with identical orientation preferences but differing retinal positions—can be used to construct translation-invariant detectors for any visual form. This is the insight underlying many networks for visual recognition including HMAX and convolutional neural nets. But, as we show here, the subsequent development of these models has been impeded by an unnecessarily limited application of the principle through which they operate. An *HW-module* with Hubel and Wiesel's proposed organization can compute invariant representations for any affine transformation. Furthermore, since it is possible to (piecewise) approximate any diffeomorphism by locally-supported affine transformations, a hierarchy of such modules is able to compute approximately invariant representations for a much wider class of non-uniform transformations.

Consider a single HW-module consisting of one C-unit and all its afferent S-units. Let the response of an S-unit to an image I be 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 S-units as template detectors. An S-unit's response is a measure of I's similarity to t. Since there are several S-units, each detecting their stored template at a different position, we introduce the translation operator $T_{\vec{x}}$ which, when applied to an image, returns its translation by $\vec{x}$. This lets us write the response of the specific S-unit which detects the presence of template t at position $\vec{x}$ as $\langle I, T_{\vec{x}}t\rangle$.

Then, introducing a nonlinear function η, the response r_C of the C-unit (equivalently: the output of the HW-module) is given by

$$r_C(I) = \sum_i \eta(\langle I, T_{\vec{x}_i} t \rangle) \tag{1.1}$$

where the sum is over all the S-units in the module. The region of space covered by a module's S-units is called the HW-module's *pooling domain* and the C-unit is said to pool the responses of its afferent S-units. This is a standard formulation (though possibly written in unfamiliar notation).

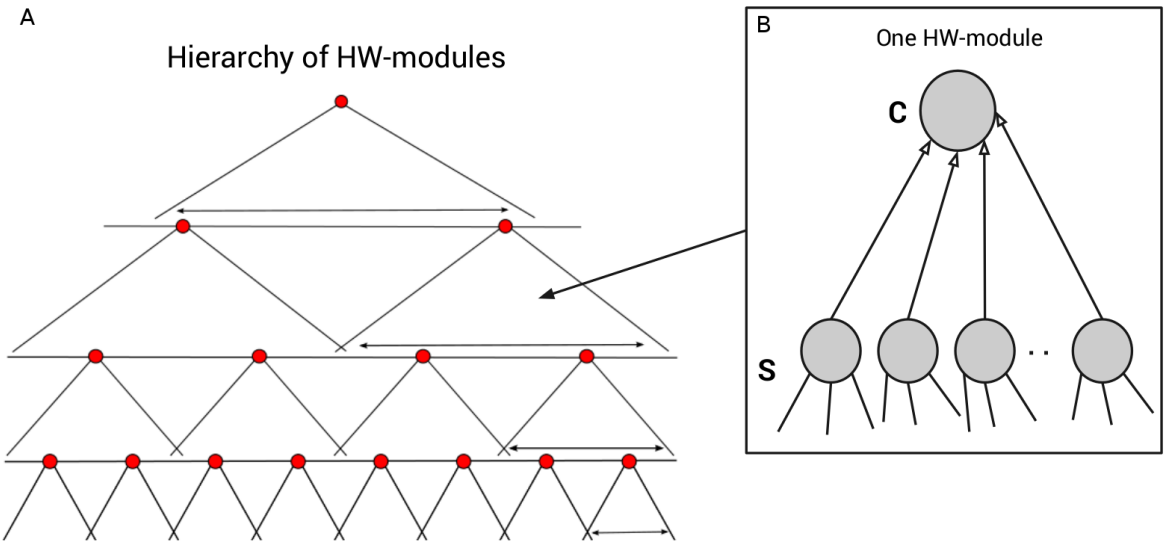

Figure 1-3: ***A.*** *A hierarchical architecture built from HW-modules. Each red circle represents the signature element computed by the associated module. Double arrows represent complex cell pooling domains, i.e., receptive fields in the case of translation. The vector computed at the top of the hierarchy consists of invariant features for the whole image.* ***B.*** *One HW-module. A set of S-units connect to one C-unit. Thanks to Fabio Anselmi for contributing the left side of this figure.*

The usual examples of HW-modules are intended as models of V1 so they generally use oriented edge or Gabor templates. However, as illustrated in figure 1-3, the principle by which HW-modules convert position-sensitive inputs into position-invariant outputs is exactly the same when other patterns are considered. Let $P = [-b, b]$ indicate a one-dimensional pooling domain. Consider a "small" pattern I. We use the

adjective "small" to give intuition for the important *localization condition* which states that the support of $\langle I, T_x t \rangle$ must be a compact subset $[-a, a]$ of P(see [1]). Write $T_{\overline{x}} I$ to indicate the image obtained after translating I by $\overline{x}$, and assume it does not fall out of the pooling domain, i.e., $|\overline{x} + a| \leq b$. Then, applying (1.1), the output of the HW-module is $r_C(I) = \sum_i \eta(\langle T_{\overline{x}} I, T_{x_i} t \rangle)$. Under what conditions is r_C invariant? Equivalently, when does $r_C(T_{\overline{x}} I) = r_C(I)$?

In the next section, we consider a vector of HW-module responses and review the conditions under which it is invariant and unique to each object.

1.3.2 Pooling and template orbits

Our aim is to compute a unique *signature* for each image x that is invariant with respect to a group of transformations G. We consider the orbit $\{gx \ : \ g \in G\}$ of x under the action of the group. In this section, G is the 2D affine group so its elements correspond to translations, scalings, and in-plane rotations of the image (notice that we use g to denote both elements of G and their representations, acting on vectors). We regard two images as equivalent if they are part of the same orbit, that is, if they are transformed versions of one another ($x' = gx$ for some $g \in G$).

The orbit of an image is itself invariant with respect to the group. For example, the set of images obtained by rotating x is exactly the same as the set of images obtained by rotating gx. The orbit is also unique for each object: the set of images obtained by rotating x only intersects with the set of images obtained by rotating x' when $x' = gx$. Thus, an intuitive method of obtaining an invariant signature for an image, unique to each object, is just to check which orbit it belongs to. We can assume access to a stored set of orbits of template images τ_k; these template orbits could have been acquired by unsupervised learning—possibly by observing objects transform and associating temporally adjacent frames (e.g. [31, 35]).

The key fact enabling this approach to object recognition is this: It is not necessary to have all the template orbits beforehand. Even with a small, sampled, set of template orbits, not including the actual orbit of x, we can still compute an invariant signature.

Observe that when g is unitary $\langle gx, \tau_k \rangle = \langle x, g^{-1}\tau_k \rangle$. That is, the inner product of the transformed image with a template is the same as the inner product of the image with a transformed template. This is true regardless of whether x is in the orbit of τ_k or not. In fact, the test image need not resemble any of the templates (see [5–7, 36]).

Consider $g_t\tau_k$ to be a realization of a random variable. For a set $\{g_t\tau_k, \quad : \quad t = 1, ..., T\}$ of images sampled from the orbit of the template τ_k, the distribution of $\langle x, g_t\tau_k \rangle$ is invariant and unique to each object. See [36] for a proof of this fact in the case that G is the group of 2D affine transformations.

Thus, the empirical distribution of the inner products $\langle x, g_t\tau_k \rangle$ is an estimate of an invariant. Following [36], we can use the empirical distribution function (CDF) as the signature:

$$\mu_n^k(x) = \frac{1}{T} \sum_{t=1}^{T} \sigma(\langle x, g_t\tau_k \rangle + n\Delta) \tag{1.2}$$

Since the distribution of the $\langle x, g_t\tau_k \rangle$ is invariant, we have many choices of possible signatures. Most notably, we can choose any of its statistical moments and these may also be invariant—or nearly so—in order to be discriminative and "invariant for a task" it only need be the case that for each k, the distributions of the $\langle x, g_t\tau_k \rangle$ have different moments. It turns out that many different convolutional networks can be understood in this framework[1]. The differences between them correspond to different choices of 1. the set of template orbits (which group), 2. the inner product (more generally, we consider the *template response function* $\Delta_{g\tau_k}(\cdot) := f(\langle \cdot, g_t\tau_k \rangle)$, for a possibly non-linear function f—see [36]) and 3. the moment used for the signature. For example, a simple neural-networks-style convolutional net with one convolutional layer and one subsampling layer (no bias term) is obtained by choosing G =translations and $\mu^k(x)$ =mean$(\cdot)$. The k-th filter is the template τ_k. The network's nonlinearity could be captured by choosing $\Delta_{g\tau_k}(x) = \tanh(x \cdot g\tau_k)$; note the similarity to Eq. (1.2). Similar descriptions could be given for modern convolutional nets, e.g. [5, 37, 38]. It is also possible to capture HMAX [3, 4] and related models (e.g. [27]) with this framework. The "simple cells" compute normalized dot products or Gaussian radial basis functions of their inputs with

[1] The computation can be made hierarchical by using the signature as the input to a subsequent layer.

stored templates and "complex cells" compute, for example, $\mu^k(x) = \max(\cdot)$. The templates are normally obtained by translation or scaling of a set of fixed patterns, often Gabor functions at the first layer and patches of natural images in subsequent layers.

The templates-and-signatures approach to recognition permits many seemingly-different convolutional networks (e.g. ConvNets and HMAX) to be understood in a common framework. We have argued here that the recent strong performance of convolutional networks across a variety of tasks (e.g., [37–39]) is explained because all these problems share a common computational crux: the need to achieve representations that are invariant to identity-preserving transformations.

Discussion

This framework for understanding invariance in object recognition cuts across levels of analysis [2]. On the level of biological implementation, its explanations of ventral stream receptive field properties are in accord with the available data. For example, it predicts Gabor tuning in V1 [40]. It can also be applied beyond V1. In that case its predictions may motivate new experiments which could potentially falsify the theory. On the computational level, it gives a unified account of *why* a range of seemingly different models have recently achieved impressive results on recognition tasks. HMAX [3], and most other convolutional networks, including [37], are among those explained. On the algorithmic level, it motivates the consideration of a new class of models which includes the previous models as special cases. We believe that the algorithm implemented by the ventral stream is in this class.

1.4 Specific contributions

1.4.1 Concerning invariance

1. An empirical measure of transformation invariance (ADT) that can be used to compare results across modeling, psychophysics, physiology, and neuroimaging experiments. —Chapter 2.

2. Demonstration that HW-modules can be approximately invariant to face viewpoint changes, and empirical investigations of the conditions under which this approach is effective. Demonstration that rotation in depth, and several other transformations (e.g., illumination and body pose) are class-specific. The need to compute invariant representations for class-specific transformations may explain the modular architecture of the anterior ventral stream i.e., why the processing of faces (and a few other classes) is separated from the circuitry processing other objects. —Chapter 4.

3. One requirement for HW-modules to be approximately invariant to non-affine transformations is for the set of templates and test objects to be nice. We developed a quantitative index of niceness that can be measured from a set of videos of objects undergoing the transformation. Among the objects we tested, faces were the nicest with respect to viewpoint changes (3D rotation in depth), and the wire objects used in [25] (and elsewhere) were the least nice. This implies that face templates can be used to recognize new faces despite depth rotations (approximately). However, templates from the wire objects are not helpful for recognizing new wire objects invariantly to depth rotations. This provides an explanation for previous psychophysics results from other groups (e.g. [41, 42]). —Chapter 4.

4. We created a suite of synthetic face datasets—which we called "Subtasks" of Unconstrained Face Recognition (SUFR)—for testing face recognition models and computer vision systems. Each dataset isolates a specific transformation. We

tested many frequently used computer vision algorithms on the full suite of subtasks as well as several "unconstrained" natural image datasets: Labeled Faces in the Wild (LFW) and a new one we gathered (SUFR-W). Qianli Liao and I contributed equally to this work. —Appendix A

1.4.2 Concerning learning

1. A Foldiak-like temporal adjacency based learning rule for HMAX which we used to model Li & DiCarlo's "invariance disruption" experiments [9, 10, 33]. Leyla Isik and I contributed equally to this work. —Chapter 3.

2. A model of macaque face patch properties using an unsupervised learning rule for HW-modules—related to Oja's and Foldiak's rules. This model explains the mysterious finding of mirror-symmetric orientation tuning curves in the face patch AL as an essential intermediate step toward the computation of an approximately view-invariant representation in patch AM. In other work [40], we (primarily Jim Mutch, Andrea Tacchetti, and Tomaso Poggio) showed that the same learning rule with a different set of parameters can also explain Gabor-like tuning in primary visual cortex. —Chapter 5.

Chapter 2

Measuring invariance

In this chapter, we numerically investigate the invariance of properties of HMAX, one architecture which is consistent with the new theory. Toward that end, we propose an operational definition of the *invariance range* (with respect to a transformation). This definition is general and could be employed in a wide variety of physiology and computational modeling studies. We find that surprisingly small numbers of templates—that need not resemble the test images—are sufficient to achieve good performance on translation and scaling invariant recognition tasks. This chapter is an updated version of a technical report we published in 2010 [7]. We also published some of this material in [8].

2.1 Introduction

This paper considers the case of initial translation invariance for novel objects. [43] showed that subjects could distinguish novel animal-like stimuli with no drop in accuracy despite shifts away from the training location of (up to) 8 degrees of visual angle. Similarly, [44] found that visual priming effects transfer between hemifields. Invariant recognition in primates depends on representations in inferotemporal visual (IT) cortex. In the IT physiology literature, smaller invariance ranges have often been reported. Typical single-unit experiments report $\sim \pm 2^\circ$ of translation invariance and ± 1 octave of scale [25, 26, 45]. Additionally, decoding analyses show that populations of IT cells

support invariant classification of viewed objects presented at shifts of up to 4 degrees from the trained location [18].

It is possible that the larger initial invariance ranges reported in some psychophysics experiments arise due to the use of stimuli that were too similar to familiar objects. Analogous translation-invariant recognition experiments using random dot pattern stimuli showed that discriminability declines when objects are presented as little as 2 degrees away from the trained location [46–49] in accord with the physiological measurements of receptive field sizes of cells in IT cortex. This paper is primarily concerned with models of this robust finding of translation invariance for novel objects in the central 4 degrees of the visual field. This is not a trivially small region over which to study invariance; the receptive field of an IT cell covering the central 4 degrees is indirectly fed by $\sim$ 125,000 cones on the retina [50]. How the ventral stream builds such invariant responses and how to replicate this process with computational models is an interesting scientific problem.

Standard models of the ventral stream by Fukushima and others [3, 27, 51, 52] attempt to replicate the process by which large IT receptive fields supporting invariant recognition are built from position-specific responses of photoreceptors (regarded as pixels in this case). In these models, units in layers corresponding to inferotemporal cortex (IT) achieve invariance due to the wiring of the model[1] —sometimes described as a convolutional architecture[2] [28].

Networks with this architecture achieve invariance using receptive fields that pool together the responses to the same set of templates at many different locations. This operation is repeated in each successively higher layer until, in the top layer, unit responses are invariant across a large region. In the case of the HMAX model [4], the templates are derived from previously-seen image fragments; a unit's response is a measure of the similarity between the image patch it's seeing now and its optimal

[1] See [53], section 2.2 for conjectures on how the wiring could be learned during development and during visual experience.

[2] The choice of a convolutional architecture corresponds to the assumption that photoreceptor density is constant over the relevant region (the central 4 degrees in this case)—an assumption that is not valid for real retinas [50]. Models that assume a convolutional architecture can be seen as describing an "upper bound" for invariance (c.f. [54] and references therein). We study convolutional architectures because their simplicity allows other aspects of the invariant recognition problem to be highlighted.

stimulus—the remembered image fragment. In the top layer, each unit will respond invariantly to its optimal stimulus no matter where it appears. The combined activities of the top-layer units are the network's representation of an object. This representation is a *signature* for the object which can be used as input by a classifier for recognition; it inherits its invariance from that of its component units.

A recent theory of learning invariant representations for object recognition [55] describes how an HMAX-like [3, 4] recognition architecture could be learned during development in an unsupervised way. In such an architecture the responses of top layer cells are invariant under affine transformations of the input image. The responses of these cells can be used as a signature for recognition—forming a vector that can be fed into a classifier. However, an invariant signature is not automatically useful for recognition. The argument in [55] guarantees that two images that are related by an affine transformation will always evoke the same signature. The converse does not need to be true. There may be a substantial numbers of false alarms—confusing images evoked by dissimilar objects[3]. This problem seems especially worrisome when small numbers of templates that do not resemble the recognition target are used. In this paper we report an empirical investigation of the extent to which invariant recognition performance is affected by the number of templates and by their similarity to the test images. To that end, we first give an operational definition of a neural population's invariance range with respect to a transformation. This definition is general and allows suitable specializations for a variety of perceptual tasks. We then show that, for a given acceptable false alarm rate, a classifier operating on [55]'s signature vector is invariant. Moreover, surprisingly small numbers of templates suffice to achieve impressively small false alarm rates—even if the templates are very dissimilar from the test images[4].

[3] But see more recent accounts of the theory, we know a lot more about the uniqueness of signatures than we did when this was written [1, 34].

[4] Another of our forthcoming papers describes more extensive empirical tests along tests along the same lines as those described here [56]. Even more recently we developed a whole set of tests, with synthetic data, that can be used systematically to measure invariance for subordinate-level face recognition tasks. The latter report is included as an appendix to this thesis A, and [57]

2.2 An operational definition of invariance

We are mostly concerned with the problem of recognizing a target object when it is presented again at a later time, usually after undergoing a transformation. In our specific problem, an image of a *target* object has already been presented. The task is to tell whenever a newly presented *test* image represents the target object. Test images may be instances of the target object under various transformations or they may be images of entirely new objects called *distractors*.

To study and measure invariance, we formulate an operational definition of invariance range in terms of classification accuracy. To do this, we first define a quantity called *Accuracy Despite Transformation* (ADT). This quantity is related to the discriminability measure (d') used in psychophysics [58]. In this context, we use ADT in order to highlight the dependence on a transformation. The quantity of interest here, the *Invariance Range*, is defined in terms of ADT.

To start, we describe a procedure for measuring translation invariance. This procedure could be used either in physiology experiments with neural data or computational simulations with computed features.

1. Choose a disk of radius r and a universe U of target and distractor objects.

2. Train a classifier C with the target at the center of the disk.

3. Test the classifier on the case where all target and distractor objects can appear anywhere on the disk with their locations drawn from a uniform distribution. Let X denote the set of images of the objects in U at all these locations.

4. We will define the *Accuracy Despite Transformation* as a summary statistic describing the classifier's performance on this test and indicate it as $ADT_{C,X}(r)$.

5. Choose a threshold θ. Increase the radius of the disk until the Accuracy Despite Transformation falls below θ. The *Invariance range* I is the largest r for which $ADT_{C,X}(r) > \theta$.

More generally, we can define the Accuracy Despite Transformation and Invariance Range for any geometric transformation. Let X denote the set of all images of targets and distractors. For a test image $x \in X$ there are two possibilities:

$$\begin{aligned} y = -1 &: x \text{ contains a distractor object} \\ y = 1 &: x \text{ contains the target object} \end{aligned}$$

The problem is described by the joint probability distribution $p(x, y)$ over images and labels.

We consider the classifier given by a function $C : X \to \mathbb{R}$ and a decision criterion η. Any choice of classifier partitions the set of images into accepted and rejected subsets: $X = X_A^\eta \cup X_R^\eta$.

$$\begin{aligned} X_A^\eta &= \{x : C(x) \geq \eta\} \\ X_R^\eta &= \{x : C(x) < \eta\} \end{aligned}$$

In an experimental setting, the classifier may refer to any decision-maker. For example, $C(x)$ could be a human observer's *familiarity* with image x, upon which the decision of "same" (target) or "different" (distractor) will be based. This is the interpretation normally taken in psychophysics and signal detection theory. Our approach is more general and could also apply to situations where $C(x)$ is interpreted as, for instance, the membrane potential of a downstream neuron or the test statistic from a machine learning classifier operating on data.

Next we define the classifier's true positive and false positive rates in the usual way:

$$TP(\eta) := \int_X P(y = 1, C(x) \geq \eta : x) p(x) dx = \text{True positive rate} \tag{2.1}$$

$$FP(\eta) := \int_X P(y = -1, C(x) \geq \eta : x) p(x) dx = \text{False positive rate} \tag{2.2}$$

Varying η generates the operating characteristic (ROC) curve:

$$ROC(\eta) = [FP(\eta), TP(\eta)] \tag{2.3}$$

Both $TP(\eta)$ and $FP(\eta)$ are defined as integrals of probability distributions. Thus, all values of $ROC(\eta)$ will fall on the unit square: $(0 \leq TP(\eta) \leq 1)$ and $(0 \leq FP(\eta) \leq 1)$. Let $\overline{ROC}(z)$ denote the ROC curve viewed as a function of the false positive rate.

We propose to use the area under the ROC curve (AUC) as a bias-free summary statistic for the definition of Accuracy under Transformation (ADT).

Definition: Accuracy Despite Transformation *For X a set of images with labels y and $P(x, y)$ the joint distribution over images and labels and $C(x)$ a classifier partitioning X according to a parameter η. Let $TP(\eta), FP(\eta)$ and the ROC curve be defined as above. The Accuracy Despite Transformation $ADT_{C,X}$ is the area under the ROC curve:*

$$ADT_{C,X} = \int_0^1 \overline{ROC}(z)dz \tag{2.4}$$

It is simple to extend this operational definition to study parametrized transformations such as translation, scaling and rotation. This is done by defining a sequence of image sets depicting the objects under wider ranges of transformation parameters. Let X be the union of a sequence of sets of images X_r ordered by inclusion.

$$X = \bigcup_r X_r \text{ with } X_r \subset X_{r'} \qquad \forall (r, r') \quad \text{ with } r' > r \tag{2.5}$$

Then we can compute the corresponding ADT_{C,X_r} for each index r. Intuitively, increasing r corresponds to increasing the "difficulty" of the task; thus ADT_{C,X_r} will be a non-increasing function of r.

As an example, translation invariance can be characterized by letting X_r contain all the images of target and distractor objects at each position in a circle of radius r. Subsequent sets $X_{r'}$ contain objects at each position within radius $r' > r$ thus $X_r \subset X_{r'}$. For a fixed classifier, we can write the Accuracy Despite Transformation as a function of the radius of the circle over which objects may appear: $ADT(r) :=$

ADT_{C,X_r}.

Finally, in order to study invariance, it is essential to separate relevant from irrelevant dimensions of stimulus variation [59, 60]. For example, the shape of the object could be the relevant dimension while its position in space may be the irrelevant dimension. The notion of Accuracy Despite Transformation extends to capture this case as well. To illustrate, we consider the case of two parametrized dimensions. Let r, s index the union $X = \bigcup_{r,s} X_{r,s}$. So for each pair r, s there is a corresponding subset of X. We require that:

$$X_{r,s} \subset X_{r',s} \text{ and } X_{r,s} \subset X_{r,s'} \qquad \forall(r, r', s, s') \qquad \text{with } r' > r \text{ and } s' > s \tag{2.6}$$

Accuracy Despite Transformation may be computed for each pair of indices r, s using equation 2.4. If one of the stimulus dimensions is not continuous then this situation is accommodated by letting either r or s take on only a few discrete values.

We can define the *invariance range* by picking a threshold level θ of Accuracy Despite Transformation and determining the maximal region size r for which $ADT(r)$ remains above θ. The same procedure could be employed when varying any aspect of the classifier (e.g. number of training examples) and determining the corresponding Accuracy Despite Transformation for each.

Definition: Invariance range *Let X_r be a sequence of sets of images ordered by inclusion, C_r an ordered sequence of classification functions and $ADT(r)$ be the classification accuracy obtained by using $C_r(x)$ to partition X_r. Let θ be a threshold value. Then the invariance range I is the maximal r for which $ADT(r) > \theta$.*

$$I = \begin{cases} \infty & ADT(r) > \theta \quad \forall r \\ \max\{r : ADT(r) > \theta\} & \text{otherwise} \end{cases} \tag{2.7}$$

In practical situations, ∞-invariance range implies good selectivity only over the entire range of transformation parameters under consideration. In these situations we will typically say “invariance range is the size of the entire visual field” or “invariance

range is the size of a downstream cell's receptive field".

Remark: The definition of invariance range given here is general. For certain choices of classifiers it can be chosen to be equivalent to other operational definitions of invariance in the literature. For example, many single-unit studies measure a quantity known as *rank-order invariance* [25, 26, 61]. In the case of translation invariance, these studies measure the rank order of the responses evoked by a set of test objects. The rankings are measured at a range of positions and a cell is said to be more invariant if the rank order of the responses to the test objects is maintained over a larger region. In the language we developed here, this procedure corresponds to choosing a nearest neighbor classifier with a particular metric for the computation of ADT[5]. The proposed definition is useful in that it focusing the discussion on the essential choices being made in any experiment that attempts to measure invariance: the choice of the universe of target and distractor objects and the choice of classifier.

Remark: If the invariance range is as large as the visual field, then we can say that there is no "selectivity-invariance trade-off". In this case objects could appear under arbitrarily extreme transformations with $ADT(r)$ never declining below θ. Conversely, a finite invariance range indicates the existence of a selectivity-invariance trade-off. A small invariance range indicates a stronger trade-off, i.e. more accuracy is lost for smaller increases in allowable transformations. We can compare selectivity-invariance trade-offs across transformations (e.g. translation and scaling versus 3D rotation or illumination) or tasks (novel versus familiar objects) by comparing their associated invariance ranges.

[5]Many studies that measure rank-order invariance calculate the response rankings separately for each position. While our definition is general enough to encompass this case, we would like to point out that a measurement which corresponds better with the intuitive notion of invariance involves measuring ADT over regions of increasing size (as described in the discussion of equation 2.5). This also allows the Invariance range I to be interpreted most straightforwardly as the last task for which performance is above a threshold where each subsequent task allows a greater range of distortion in the target objects.

2.3 Measuring Accuracy Despite Transformation

2.3.1 Physiology

We can measure $ADT(r)$ in physiology experiments. A typical experiment consists of an animal viewing stimuli, either passively or while engaged in a task, at the same time as the experimenter records neural data. The neural data could consist of any of the various electrophysiological or neuroimaging methods. In the case of a single-unit electrophysiology experiment, the data consists of the evoked firing rates of a collection of cells in response to a stimulus.

Assume we have recorded from n cells while presenting images of target $y = 1$ and distractor objects $y = -1$. Define a classifier $C(x)$ on the neural responses evoked by each image. Then vary the threshold η accepting images with $C(x) > \eta$ to draw out an ROC curve. $ADT(r)$ is the area under the ROC curve for each r.

Remark: Most electrophysiology experiments will not allow simultaneous recordings from more than a few cells. In order to obtain larger populations of cells you can bring together cells recorded at different times as long as their responses were evoked by the same stimuli. These *pseudopopulations* have been discussed at length in elsewhere [18, 62].

Remark: We can also measure ADT from other kinds of neural data. Many researchers have recorded fMRI data while presenting a human or animal subject with stimuli. Replace cells with fMRI voxels and apply the same measurement process as for single-unit electrophysiology. This is a particular application of "multi-voxel pattern analysis". The same approach can be taken to analyze Magnetoencephalography (MEG) data. See [19, 63] for examples.

2.3.2 Computer Vision

Many computer object recognition systems can be described as having two basic modules. An initial feature transformation converts input images into a new representation. Then a classifier $C(x)$ operates on a set of feature-transformed versions of images to

answer questions about the images e.g. do two test images correspond to the same object? We can directly compute ADT using the classifier and the labels y_x.

Remark: The classifier must be chosen appropriately for the task being modeled. For example, a same-different task involving novel objects appearing under various transformations, e.g. [46, 49], could be modeled with a classifier trained on a single example of the target image that accepts inputs judged likely to be a transformed version of the trained image and rejects inputs judged likely to be distractors. An alternative same-different task using familiar objects which the subject had seen at all positions prior to the experiment could be modeled using a classifier involving a larger number of training images under different transformation conditions.

Remark: The area under the ROC curve for popular same-different matching tasks like Labeled Faces in the Wild [64] has a somewhat similar interpretation to ADT. However, in that case the images are considerably less constrained and it is not possible to attribute performance (or lack thereof) to how well a model tolerates (or fails to tolerate) a particular transformation. See Appendix A.2.1 for more additional results on transformation-invariant face recognition and its relationship to performance on natural image datasets.

2.3.3 Psychophysics

We can compute ADT in behavioral experiments by making a few extra assumptions. First, the subject must be engaged in a same-different task e.g. the task is to accept images that show the target object and reject images that show a distractor object. Test images may be transformed versions of either target or distractor objects.

We regard the subject's response analogously to the thresholded output of a classifier. The subject's choice of a decision criterion - called *response bias* in this context - is not controllable by the experimenter. However, we can still estimate the area under the ROC curve without explicit access to the threshold as long as we assume that the underlying distributions P_N and P_P are both Gaussian. This is the standard assumption of signal detection theory [58]. In this case ADT is related to the standard psychophysical

measure of discriminability d' by the following:

$$\text{ADT} = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{d'}{2}\right),$$

where erf$()$ denotes the error function:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$$

and $d' = Z(TP) - Z(FP)$ where $Z()$ denotes a Z-score. See [65] for a simple derivation of this relationship.

2.3.4 Invariant signatures for classification

In the setting we have developed here, the natural recognition architecture, for both the ventral stream, as well as for computer vision systems, is one in which a classifier is trained and tested on the signatures of its inputs. In this section we discuss the Invariance range and Accuracy Despite Transformation achieved by such systems.

Consider a classification function $C_\Sigma : \Sigma_X \to \mathbb{R}$ operating on the signatures of images. We are particularly interested in the case of novel objects—just one training example x_0. A natural choice of classifier for this case is:

$$C_\Sigma(\Sigma_x) = e^{-\|\Sigma_x - \Sigma_{x_0}\|_2^2} \tag{2.8}$$

As in section 2.2 we can define a sequence of progressively more difficult classification problems by testing on sets containing a wider range of transformations of the original images. Then, as above, we can define the ADT_{C,X_r} for each r.

$$X_r = \bigcup_p T_p(X_0) \quad \text{for } 0 \le p \le r \tag{2.9}$$

If T_r is a family of affine transformations then we know that $\Sigma_x = \Sigma_{T_r x}$ for all r (see 1.3.2). Since the signature is unchanged by T_r, it follows that the classification problem on X_r is exactly the same as the classification problem on X_0 for all r. Thus:

$$ADT_{C_\Sigma,X_r} = ADT_{C_\Sigma,X_0} \quad \text{for all } r \tag{2.10}$$

This means that, for affine transformations, a plot of ADT_{C_Σ,X_r} as a function of r will always be a flat line. A system that classifies images based on their encoding relative to an invariant signature achieves the same level of performance when tested on transformed test images as it does on untransformed test images. It is always possible to choose a threshold θ so that the invariance range I is ∞. This is true even if only a single example image is available for training.

2.3.5 Discriminability

We showed above that these signature-based systems can always achieve ∞-invariance range. We can then ask about how discriminative these systems are. In our language, this is a question about the Accuracy Despite Transformation. Since ADT_{C_Σ,X_r} is a constant function of r we only need to consider ADT_{C_Σ,X_0}.

It is easy to construct systems where ADT_{C_Σ,X_0} is very low. For example, consider a system with a single template that has 0 similarity to all the test images. This signature would map all images to the same point and any classifier using this representation would never achieve a performance level better than chance. The invariance range would be ∞, but the system would still be worthless. This will happen whenever the set of template images is not rich enough to accurately preserve the difference between target and distractor images.

In the remainder of this paper we describe the results of numerical simulations of translation and scaling-invariant object recognition using different sets of templates. In section 2.4.1 we show the performance of the HMAX model in a situation similar to one that is commonly tested in physiology experiments. In section 2.4.2, we investigate the role of the similarity of the templates to the test images and the number of templates used to form the signature.

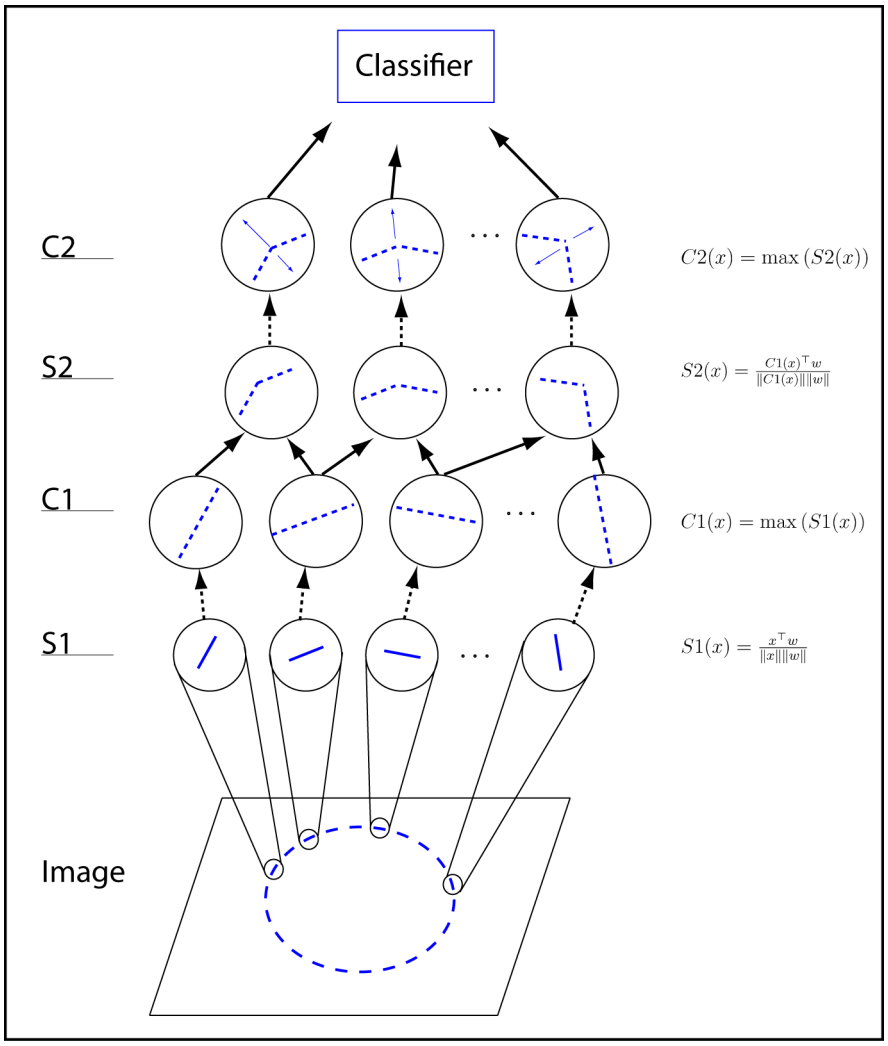

Figure 2-1: *At the first stage of processing, the S1 units compute the responses of Gabor filters (at 4 orientations) with the image's (greyscale) pixel representation. At the next step of processing, C1 units pool over a set of S1 units in a local spatial region and output the single maximum response over their inputs. Thus a C1 unit will have a preferred Gabor orientation but will respond invariantly over some changes in the stimulus' position. Each layer labeled "S" can be regarded as computing a selectivity-increasing operation (normalized dot product) while the "C" layers perform invariance-increasing operations. The weights w at layer S1 define the Gabor filters. At layer $S2$, the weights are the stored templates. In this case, we chose these weights to either be (C1-encoded) natural images or random dot patterns.*

2.4 Simulations

For numerical simulations of invariant object recognition, we used a simple version of the HMAX model from [52][6]. This 4-layer model converts images into signature vectors

[6]Matlab code to run this model is available online at http://cbcl.mit.edu/jmutch/cns/

via a series of processing stages referred to as layers. In order, the layers of the model were: S1 $\rightarrow$ C1 $\rightarrow$ S2 $\rightarrow$ C2. Unlike most implementations of the HMAX model, we used only a single scale.

At the first stage of processing, the S1 units compute the responses of Gabor filters (at 4 orientations) with the image's (greyscale) pixel representation (256 x 256 pixels). At the next step of processing, C1 units pool over a set of S1 units in a local spatial region and output the single maximum response over their inputs. Thus a C1 unit will have a preferred Gabor orientation but will respond invariantly over some changes in the stimulus's position. The S2 units signal the similarity of their input to a stored template by computing a normalized dot product of the input with the template. The model has a "convolutional architecture" because there is an S2 cell at every position, tuned to every template's appearance at that position. The S2 stored templates are either patches of natural images or random dot patterns; see the figure captions for the choice of templates in each simulation.

The S2 units compute the following function of their inputs $x = (x_1, ...x_n)$

$$r = \exp\left(-\frac{1}{2\sigma}\sum_{j=1}^{n}(w_j - x_j)^2\right) \tag{2.11}$$

All of the simulations were performed as follows. The classifier ranked all images by their correlation to a single "trained" image. The trained image was always a single image containing the target at the center of the transformation range.

All the AUC (equivalently: ADT) values reported here are averages over several simulations. Each dataset contained a number N of objects. We chose each in turn to be the target object and used the remaining $N-1$ objects as distractors. The reported AUC values are the means over all N simulations.

For all the translation invariance experiments, targets and distractors appeared only on an interval of length $2r$ as opposed to the entire disk of radius r. This was done for computational efficiency reasons. We also repeated a subset of these simulations using the full disk and obtained the same results.

The templates, τ_n (preferred featured of S2 units) were patches of images. These

were chosen differently for each simulation; details are in the captions of the relevant figures.

2.4.1 Translation and scaling-invariant recognition

Figure 2-2A shows the model is translation invariant. It has $\infty-$invariance range. Classification is based on a signature with templates tuned to patches of natural images (2000 templates). The target class consists of an image of a single face appearing at various positions in the visual field. All the variation in the target class is due to translations. The classifier measures the distance between the signature vector of the reference image and any input. In accord with the analysis above, the performance is perfect ($ADT = 1$), and unaffected by translation (red curve). The blue curve shows the results from a classifier trained in the same way, operating on a lower layer of the HMAX model. In this layer, the model does not aggregate all the transformed versions of each template so recognition performance is not invariant.

The simulation shown in figure 2-2B complicates the perceptual task by allowing greater variability in the target class. In this case, the target class consisted of 10 images of the same face acquired under slight variations in pose and facial expression. The classifier still only sees one positive example (it measures the input's similarity to a single reference image), however, it now has to correctly generalize over variations in pose and expression in addition to discounting translation. The results show that translation does not "interact" with these other sources of variability in image appearance. The same performance level is obtained no matter how extreme the translations. The invariance range is still ∞ (i.e. the size of the readout cell's receptive field) for this more realistic perceptual task.

2.4.2 Accurate recognition with small numbers of random dot templates

Previous models have assumed that that templates only enable recognition of novel objects that resemble the templates or are built from them (as for instance Ullman's

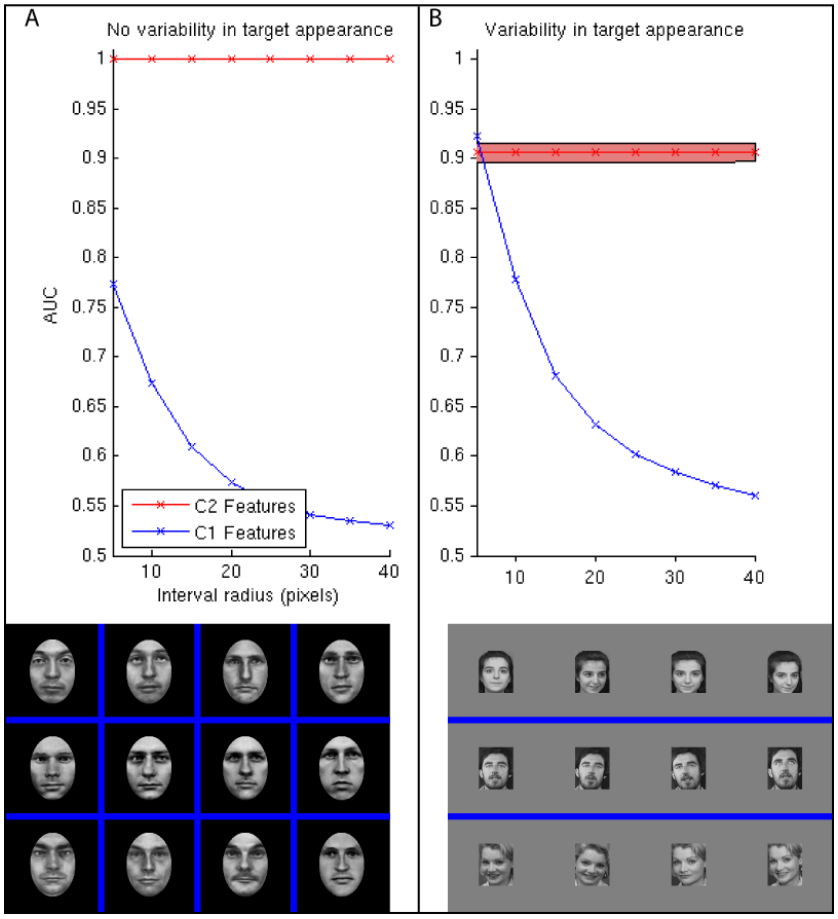

Figure 2-2: *Accuracy under Translation over intervals of increasing size. The size of the interval over which targets and distractors could appear is plotted as the abscissa with the corresponding ADT as the ordinate. For these simulations there were 2000 C2 layer cells with patches randomly sampled from natural images. Panel A: The classifier computes the correlation from the representation of a target face presented at the center to the representation of an input face presented at variable locations. The targets and distractors are faces modified from the Max Planck Institute face database [66] The images are 256x256 pixels and the faces are 120 pixels across. Panel B: The classifier still computes the correlation from the representation of a target face presented at the center to the representation of an input face presented with variable location. However, now the positive class contains additional images of the same person (slightly variable pose and facial expression). A perfect response would rank the entire positive class as more similar to the single "trained" example than any members of the negative class. The images used in this simulation were modified from the ORL face dataset, available from AT&T laboratories, Cambridge http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. These images were 256x256 pixels and the translated face was 100 pixels across. The error bars show +/-1 standard deviation over multiple runs of the simulation using different templates.*

fragments are [67]). Our results, instead, support the alternative hypothesis: most objects can be recognized using an invariant signature based on computing the similarity

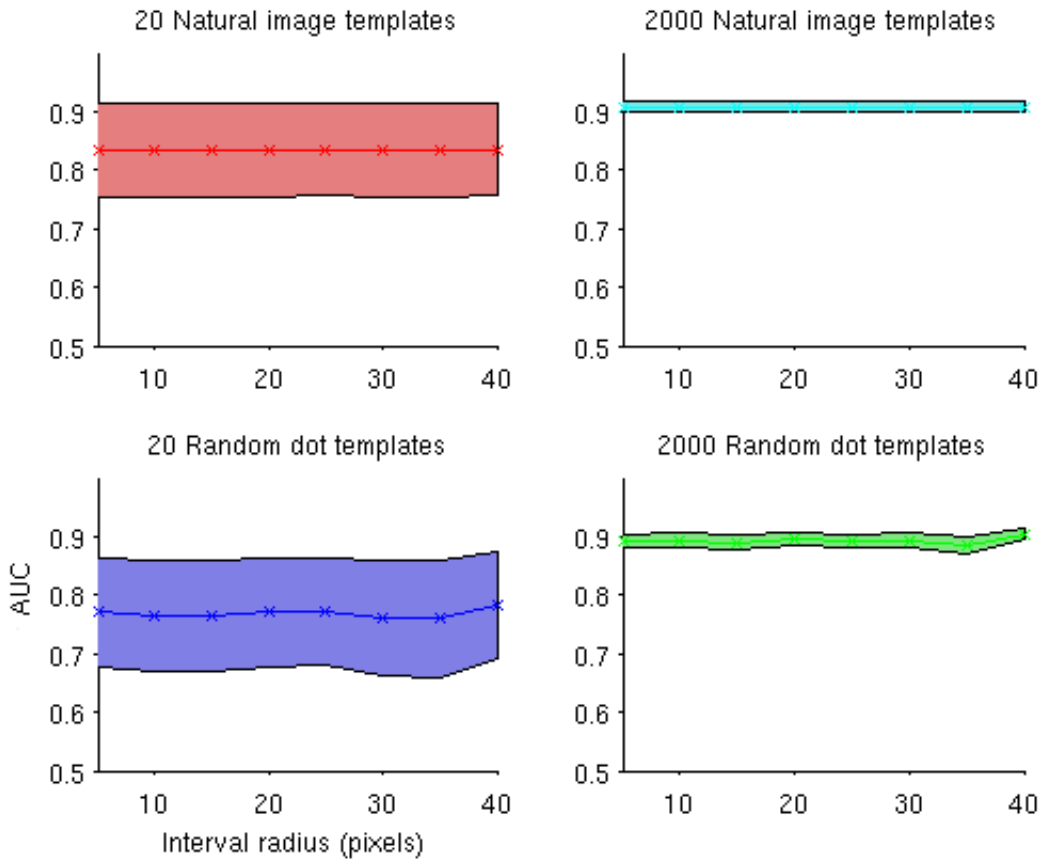

Figure 2-3: *AUC (ADT) for recognizing targets versus distractors appearing anywhere within an interval of a particular radius (using C2-layer features only). Upper left: 2000 templates were extracted from natural images. Upper right: 2000 templates extracted from random dot patterns. Bottom row: 20 templates extracted from natural images and random dot patterns respectively. We build invariant signatures by sampling randomly chosen patches from the template set of images. Error bars display +/- one standard deviation. The test images here were the same as those used in figure 2-2B.*

to essentially any set of templates.

As evident from figure 2-3 (bottom right panel), even invariant templates extracted from highly unnatural objects (random dot patterns) are sufficient to support invariant face identification (same task as in figure 2-2B). The left panels display the results from simulations with very small numbers of templates (20) drawn from images of either random dot patterns (blue curve) or natural images (red curve). Surprisingly, these results show that the ∞-invariance range for affine transformations is maintained despite using these very small numbers of templates. Figure 2-4 shows classification accuracy as a

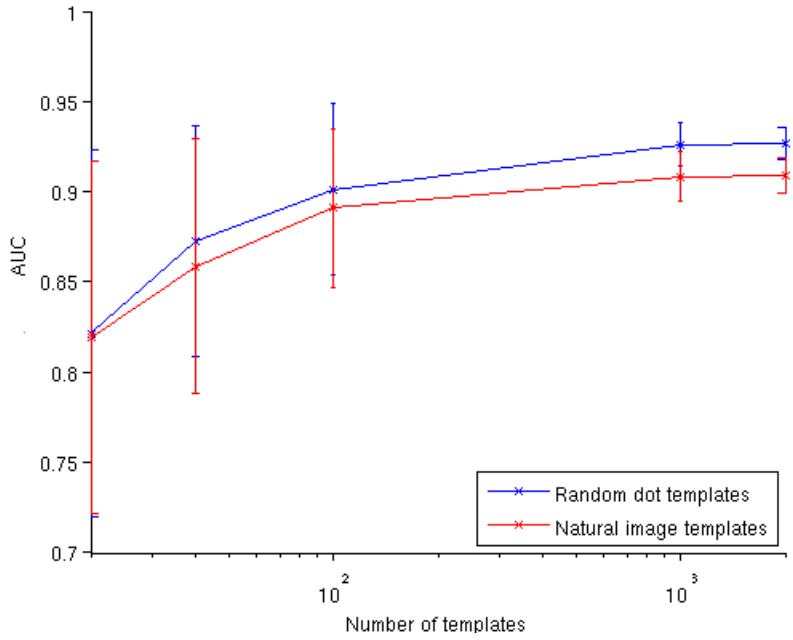

Figure 2-4: *AUC (ADT) for the face identification task as a function of the number of templates used to build the model. Here invariant templates were derived from random patches of 20 natural images (red curve) or 20 random dot patterns (blue curve). Restarting the simulation and choosing different random patches gave us a measure of the variability. Error bars are +/- one standard deviation. The test images here were the same as those used in figure 2-2B.*

function of the number and type of templates.

2.5 Discussion

Geometric transformations distort an object's appearance while preserving its identity. Recognizing objects by the images they evoke requires a system that discounts such transformations. Models of object recognition, and the ventral stream, can be interpreted as computing a signature vector that measures an input image's similarity to a set of previously encountered template images. As long as the transformations to be discounted are all affine, the signature can, in principle, be completely invariant. A classifier operating on these invariant signatures can be used to recognize novel objects from a single example image.

Recognizing objects using invariant signatures could lead to false alarms. However,

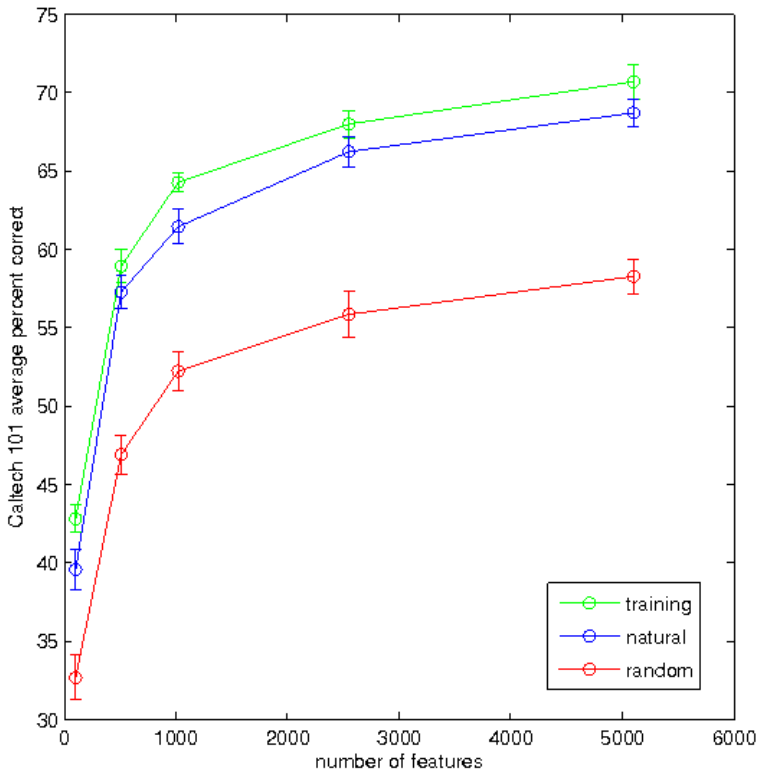

Figure 2-5: *Average percent correct on the Caltech 101 dataset [68] using a regularized least squares (RLS) classifier [69] with 30 training images per class. The templates were extracted from random patches drawn from either the training images (green curve), an unrelated set of natural images (blue curve) or random dot patterns (red curve). Each datapoint is the average accuracy from 8 runs using different training/test image splits and different randomly selected templates. The error bars show +/- one standard deviation. This simulation used the publicly available implementation of the HMAX model accompanying [4].*

increasing the number of templates decreases the false alarm rate. Computational simulations of translation invariance show that small numbers of templates suffice to achieve low false alarm rates. Moreover, the templates do not need to resemble the test images. To investigate the role of task-suitable features more generally, we tested the same random dot templates on a standard computer vision dataset for natural image categorization (figure 2-5). The surprisingly strong performance of random dot templates on this task can be attributed to the translation and scale invariance of HMAX. It seems that a substantial portion of what makes this categorization task difficult can

be attributed to training and test images appearing at different positions and scales. This result is related to other recent studies that found convolutional architectures with random templates perform well on a variety of computer vision tasks [5, 6, 70].

2.6 Acknowledgments

We would like to thank Jim Mutch for many discussions relating to this project and his contributing the Caltech101 simulation in figure 2-5. This report describes research done at the Center for Biological and Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Department of Brain and Cognitivie Sciences, and which is affiliated with the Computer Science and Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-050C-7262). Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation.

Chapter 3

Learning and disrupting invariance in visual recognition with a temporal association rule

Leyla Isik and I contributed equally to this chapter

Learning by temporal association rules such as Foldiak's trace rule is an attractive hypothesis that explains the development of invariance in visual recognition. Consistent with these rules, several recent experiments have shown that invariance can be broken at both the psychophysical and single cell levels. We show a) that temporal association learning provides appropriate invariance in models of object recognition inspired by the visual cortex, b) that we can replicate the "invariance disruption" experiments using these models with a temporal association learning rule to develop and maintain invariance, and c) that despite dramatic single cell effects, a population of cells is very robust to these disruptions. We argue that these models account for the stability of perceptual invariance despite the underlying plasticity of the system, the variability of the visual world and expected noise in the biological mechanisms.

3.1 Introduction

A single object can give rise to a wide variety of images. The pixels (or photoreceptor activations) that make up an image of an object change dramatically when the object is moved relative to its observer. Despite these large changes in sensory input, the brain's ability to recognize objects is relatively unimpeded. Temporal association methods are promising solutions to the problem of how to build computer vision systems that achieve similar feats of invariant recognition [31, 35, 51, 71–76]. These methods associate temporally adjacent views under the assumption that temporal adjacency is usually a good cue that two images are of the same object. For example, an eye movement from left to right causes an object to translate on the visual field from right to left; under such a rule, the cells activated by the presence of the object on the right will be linked with the cells activated by the presence of the object on the left. This linkage can be used to signal that the two views represent the same object—despite its change in retinal position.

Recent experimental evidence suggests that the brain may also build invariance with this method. Furthermore, the natural temporal association-based learning rule remains active even after visual development is complete [9, 10, 32, 33, 77]. This paper addresses the wiring errors that must occur with such a continually active learning rule due to regular disruptions of temporal contiguity (from lighting changes, sudden occlusions, or biological imperfections, for example.)

Experimental studies of temporal association involve putting observers in an altered visual environment where objects change identity across saccades. Cox *et al.* showed that after about an hour of exposure to an altered environment, where objects changed identity at a specific retinal position, the subjects mistook one object for another at the swapped position while preserving their ability to discriminate the same objects at other positions [9]. A subsequent physiology experiment by Li and DiCarlo using a similar paradigm showed that individual neurons in primate anterior inferotemporal cortex (AIT) change their selectivity in a position-dependent manner after less than an hour of exposure to the altered visual environment [10].

The Li and DiCarlo experiment did not include a behavioral readout, so the effects of the manipulation on the monkey's perception are not currently known; however, the apparent robustness of our visual system suggests it is highly unlikely that the monkey would really be confused between such different looking objects (e.g. a teacup and a sailboat) after such a short exposure to the altered visual environment. In contrast, the Cox *et al.* psychophysics experiment had a similar timecourse (a significant effect was present after one hour of exposure) but used much more difficult to discriminate objects ("Greebles" [78]).

In this paper, we describe a computational model of invariance learning that shows how strong effects at the single cell level, like those observed in the experiments by Li and DiCarlo do not necessarily cause confusion on the neural population level, and hence do not imply perceptual effects. Our simulations show that a population of cells is surprisingly robust to large numbers of mis-wirings due to errors of temporal association.

3.2 Simulation methods

3.2.1 Hierarchical models of object recognition

We examine temporal association learning with a class of cortical models inspired by Hubel and Wiesel's famous studies of visual cortex [21]. These models contain alternating layers of simple S-cells or feature detectors to build specificity, and complex C-cells that pool over simple cells to build invariance. [3, 4, 27]. We will focus on one particular such model, HMAX [4]. The differences between these models are likely irrelevant to the issue we are studying, and thus our results will generalize to other models in this class.

3.2.2 The HMAX model

In this model, simple (S) cells compute a measure of their input's similarity to a stored optimal feature via a Gaussian radial basis function (RBF) or a normalized dot product.

Complex (C)-cells pool over S-cells by computing the *max* response of all the S cells with which they are connected. These operations are typically repeated in a hierarchical manner, with the output of one C layer feeding into the next S layer and so on. The model used in this report had 4 layers: S1 $\rightarrow$ C1 $\rightarrow$ S2 $\rightarrow$ C2. The caption of figure 3-1 gives additional details of the model's structure.

In our implementation of the HMAX model, the response of a C2 cell – associating templates w at each position t – is given by:

$$r_w(x) = \max_t \left(\exp\left(-\frac{1}{2\sigma} \sum_{j=1}^{n} (w_{t,j} - x_j)^2 \right) \right) \tag{3.1}$$

In the hardwired model, each template w_t is replicated at all positions, thus the C2 response models the outcome of a previous temporal association learning process that associated the patterns evoked by a template at each position. The C2 responses of the hardwired model are invariant to translation [4, 7]. The remainder of this report is focused on the model with learned pooling domains. Section 3.2.3 describes the learning procedure and figure 3-2 compares the performance of the hardwired model to an HMAX model with learned C2 pooling domains.

As in [4], we typically obtain S2 templates from patches of natural images (except where noted in figure 3-3). The focus of this report is on learning the pooling domains. The choice of templates, i.e„ the learning of selectivity (as opposed to invariance) is a separate issue with a large literature of its own[1].

3.2.3 Temporal association learning

Temporal association learning rules provide a plausible way to learn transformation invariance through natural visual experience [31, 35, 51, 71**?** –75]. Objects typically move in and out of our visual field much slower than they transform due to changes in pose and position. Based on this difference in timescale we can group together cells that are tuned to the same object under different transformations.

[1] See [7] for a discussion of the impact of template-choice on HMAX results with a similar translation-invariant recognition task to the one used here

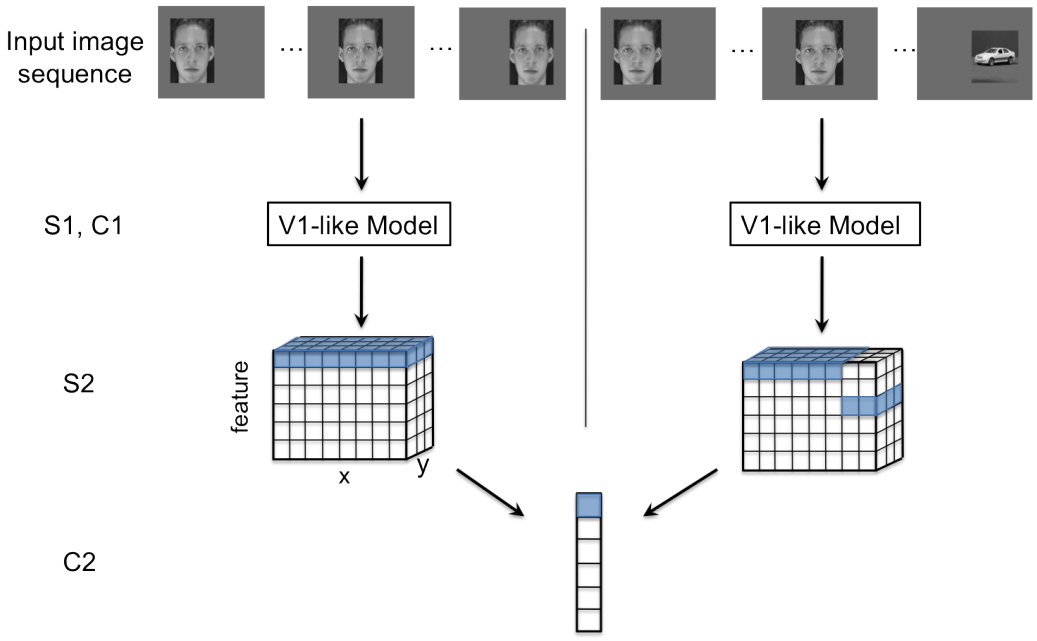

Figure 3-1: An illustration of the HMAX model with two different input image sequences: a normal translating image sequence (left), and an altered temporal image sequence (right). The model consists of four layers of alternating simple and complex cells. **S1 and C1 (V1-like model)**: The first two model layers make up a V1-like model that mimics simple and complex cells in the primary visual cortex. The first layer, S1, consists of simple orientation-tuned Gabor filters, and cells in the following layer, C1, pool (maximum function) over local regions of a given S1 feature. **S2**: The next layer, S2, performs template matching between C1 responses from an input image and the C1 responses of stored prototypes (unless otherwise noted, we use prototypes that were tuned to, C1 representations of, natural image patches). Template matching is implemented with a radial basis function (RBF) network, where the responses have a Gaussian-like dependence on the Euclidean distance between the (C1) neural representation of an input image patch and a stored prototype. The RBF response to each template is calculated at various spatial locations for the image (with half overlap). Thus the S2 response to one image (or image sequence) has three dimensions: x and y corresponding to the original image dimensions, and feature the response to each template. **C2**: Each cell in the final layer, C2, pools (maximum function) over all the S2 units to which it is connected. The S2 to C2 connections are highlighted for both the normal (left) and altered (right) image sequences. To achieve ideal transformation invariance, the C2 cell can pool over all positions for a given feature as shown with the highlighted cells.

Our model learns translation invariance from a sequence of images of continuously translating objects. During a training phase prior to each simulation, the model's S2 to C2 connections are learned by associating the patterns evoked by adjacent images in

the training sequence as shown in Figure 3-1, left.

The training phase is divided into temporal association periods. During each temporal association period the highly active S2 cells become connected to the same C2 cell. One C2 cell is learned during each association period. When modeling "standard" (undisrupted) visual experience, as in figure 3-2, each association period contains all views of a single object at each retinal position. If temporally adjacent images really depict the same object at different positions, then this procedure will group all the S2 cells that were activated by viewing the object, no matter what spatial location elicited the response. The outcome of an association period is illustrated in Figure 3-1, left. The C2 cell produced by this process pools over its connected S2 cells. The potential effect of a temporally altered image sequence is illustrated in Figure 3-1, right. This altered training will likely result in mis-wirings between the S2 and C2 neurons, which could ultimately alter the system's performance.

Learning rule

In Foldiak's original trace rule, shown in Equation 3.2, the weight of a synapse w_{ij} between an input cell x_j and output cell y_i is strengthened proportionally to the input activity and the trace or average of recent output activity at time t. The dependence of the trace on previous activity decays over time with the δ term [31].

Foldiak trace rule:

$$\Delta w_{ij}^{(t)} \propto x_j \bar{y}_i^{(t)}$$
$$\bar{y}_i^{(t)} = (1-\delta) y_i^{(t-1)} + \delta y_i^{(t)} \tag{3.2}$$

In the HMAX model, connections between S and C cells are binary. Additionally, in our training case we want to learn connections based on image sequences of a known length, and thus for simplicity should include a hard time window rather than a decaying time dependence. Thus we employed a modified trace rule that is appropriate for learning S2 to C2 connections in the HMAX model.

Modified trace rule for the HMAX model:

$$\begin{aligned} &\text{for } t \text{ in } \tau: \\ &\qquad \text{if } x_j > \theta,\ w_{ij} = 1 \\ &\qquad \text{else},\ w_{ij} = 0 \end{aligned} \tag{3.3}$$

With this learning rule, one C2 cell with index i is produced for each association period. The length of the association period is τ.

3.3 Results

3.3.1 Training for translation invariance

We model natural invariance learning with a training phase where the model learns to group different representations of a given object based on the learning rule in Equation 3.3. Through the learning rule, the model groups continuously-translating images that move across the field of view over each known association period τ. An example of a translating image sequence is shown at the top, left of Figure 3-1. During this training phase, the model learns the domain of pooling for each C2 cell.

3.3.2 Accuracy of temporal association learning

To test the performance of the HMAX model with the learning rule in Equation 3.3, we train the model with a sequence of training images. Next we compare the learned model's performance to that of the hard-wired HMAX [4] on a translation-invariant recognition task. In standard implementations of the HMAX model, the S2 to C2 connections are hard-wired, each C2 cell pools all the S2 responses for a given template globally over all spatial locations. This pooling gives the model translation invariance and mimics the outcome of an idealized temporal association process.

The task is a 20 face and 20 car identification task, where the target images are

similar (but not identical) for different translated views[2]. We collect hard-wired C2 units and C2 units learned from temporal sequences of the faces and cars. We then used a nearest neighbor classifier to compare the correlation of C2 responses for translated objects to those in a given reference position. The accuracy of the two methods (hard-wired and learned from test images) versus translation is shown in Figure 3-2. The two methods performed equally well. This confirms that the temporal associations learned from this training yield accurate invariance results.

3.3.3 Manipulating the translation invariance of a single cell

In their physiology experiments Li and DiCarlo identified AIT cells that responded preferentially to one object over another, they then performed altered temporal association training where the two objects were swapped at a given position [10]. To model these experiments we perform temporal association learning described by Equation 3.3 with a translating image of one face and one car. For this simulation, the S2 units are tuned to the same face and car images (see Figure 3-1 caption) to mimic object-selective cells that are found in AIT. Next we select a “swap position" and perform completely new, altered training with the face and car images swapped only at that position (see Figure 3-1, top right). After the altered training, we observe the response (of one C2 cell) to the two objects at the swap position and another non-swap position in the visual field that was unaltered during training.

As shown in Figure 3-3, the C2 response for the preferred object at the swap position (but not the non-swap position) is lower after training, and the C2 response to the non-preferred object is higher at the swap position. As in the physiology experiments performed by Li and DiCarlo, these results are object and position specific. Though unsurprising, this result draws a parallel between the response of a single C2 unit and the physiological response of a single cell.

[2]The invariance-training and testing datasets come from a concatenation of two datasets from: ETH80 (http://www.d2.mpi-inf.mpg.de/Datasets/ETH80) and ORL (http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html). Except when noted, the image patches used to obtain the S2 templates were obtained from a different, unrelated, collection of natural images; see [4] for details.

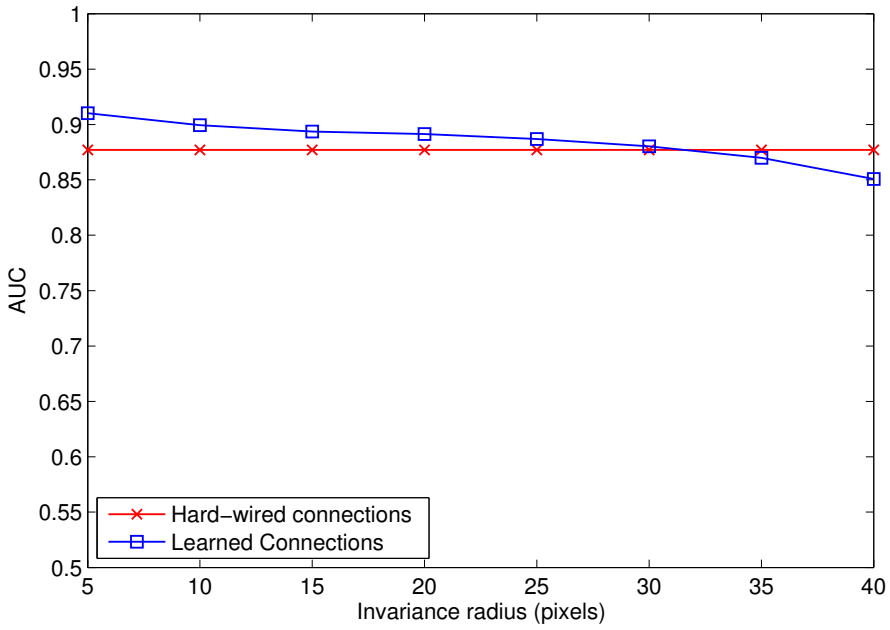

Figure 3-2: The area under the ROC curve (AUC) (ordinate) plotted for the task of classifiying (nearest neighbors) objects appearing on an interval of increasing distance from the reference position (abscissa). The model was trained and tested on separate training and testing sets, each with 20 car and 20 face images. For temporal association learning, one C2 unit is learned for each association period or training image, yielding 40 learned C2 units. One hard-wired C2 unit was learned from each natural image patch that S2 cells were tuned to, yielding 10 hard wired C2 units. Increasing the number of hard-wired features has only a marginal effect on classification accuracy. For temporal association learning, the association period τ was set to the length of each image sequence (12 frames), and the activation threshold θ was empirically set to 3.9 standard deviations above the mean activation.

3.3.4 Individual cell versus population response

In the previous section we modeled the single cell results of Li and DiCarlo, namely that translation invariant representations of objects can be disrupted by a relatively small amount of exposure to altered temporal associations. However, single cell changes do not necessarily reflect whole population or perceptual behavior and no behavioral tests were performed on the animals in this study.

A cortical model with a temporal association learning rule provides a way to model population behavior with swap exposures similar to the ones used by Li and DiCarlo [10, 33]. A C2 cell in the HMAX model can be treated as analogous to an AIT cell (as

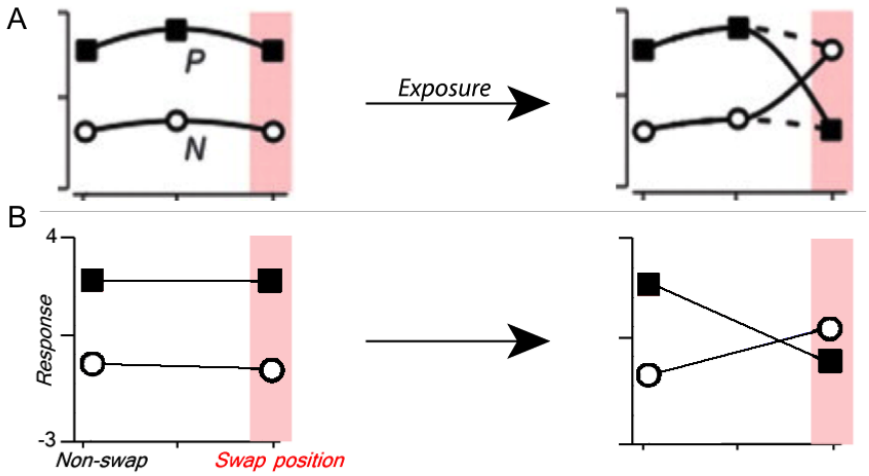

Figure 3-3: Manipulating single cell translation invariance through altered visual experience. (A) Figure from Li and DiCarlo 2008 [10] summarizing the expected results of swap exposure on a single cell. P is the response to preferred stimulus, and N is that to non-preferred stimulus. (B) The response of a C2 cell tuned to a preferred object before (left) and after (right) altered visual training where the preferred and non-preferred objects were swapped at a given position. To model the experimental paradigm used in [9, 10, 32, 33], altered training and final testing were performed on the same altered image sequence. The C2 cell's relative response (Z-score) to the preferred and non-preferred objects is shown on the ordinate, and the position (swap or non-swap) is shown on the abscissa.

tested by Li and DiCarlo), and a C2 vector as a population of these cells. We can thus apply a classifier to this cell population to obtain a model of behavior or perception.

3.3.5 Robustness of temporal association learning with a population of cells

We next model the response of a population of cells to different amounts of swap exposure, as illustrated in Figure 3-1, right. The translating image sequence with which we train the model replicates visual experience, and thus jumbling varying amounts of these training images is analogous to presenting different amounts of altered exposure to a test subject as in [10, 33]. These disruptions also model the mis-associations that may occur with temporal association learning due to sudden changes in the visual field (such as light, occlusions, etc), or other imperfections of the biological learning mechanism. During each training phase we randomly swap different face and car images in

the image sequences with a certain probability, and observe the effect on the response of a classifier to a population of C2 cells. The accuracy (AUC) versus different neural population sizes (number of C2 cells) is shown in Figure 3-4 for several amounts of altered exposure. We measured altered exposure by the probability of flipping a face and car image in the training sequence.

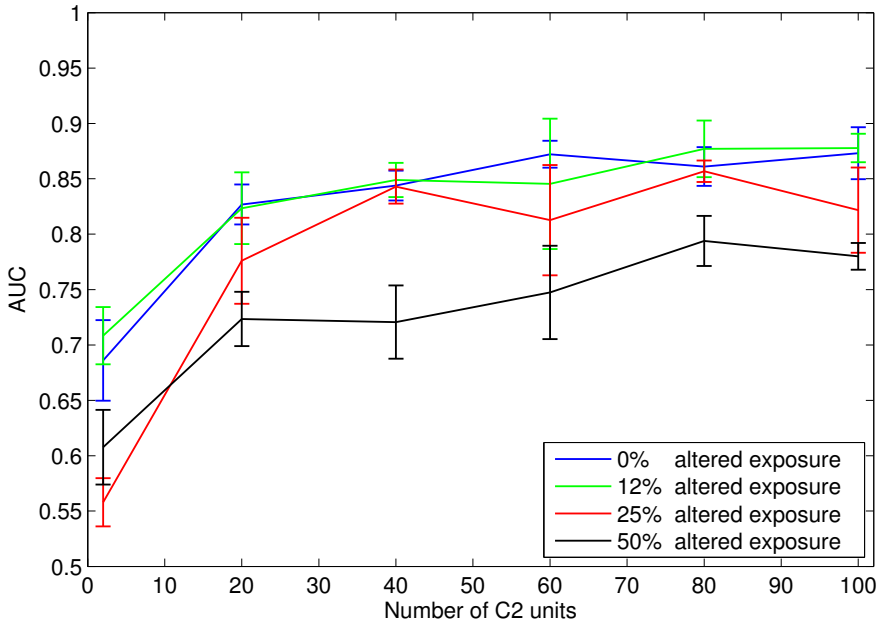

Figure 3-4: Results of a translation invariance task (+/- 40 pixels) with varying amounts of altered visual experience. To model the experimental paradigm used in [9, 10, 32, 33, 77], training and testing were performed on the same altered image sequence. The performance (AUC) on the same translation-invariant recognition task as in figure 3-2, with a nearest neighbor classifier, versus the number of C2 units. Different curves have a different amount of exposure to altered visual training as measured by the probability of swapping a car and face image during training. The error bars show +/- one standard deviation.

A small amount of exposure to altered temporal training (0.125 probability of flipping each face and car) has negligible effects, and the model under this altered training performs as well as with normal temporal training. A larger amount of exposure to altered temporal training (0.25 image flip probability) is not significantly different than perfect temporal training, especially if the neural population is large enough. With enough C2 cells, each of which is learned from a temporal training sequence, the effects of small

amounts of jumbling in training images are insignificant. Even with half altered exposure (0.5 image flip probability), if there are enough C2 cells then classification performance is still reasonable. This is likely because with similar training (multiple translating faces or cars), redundant C2 cells are formed, creating robustness to association errors that occurred during altered training. Similar redundancies are likely to occur in natural vision. This indicates that in natural learning mis-wirings do not have a strong effect on learning translation invariance, particularly with familiar objects or tasks.

3.4 Discussion

We use a cortical model inspired by Hubel and Wiesel [21], where translation invariance is learned through a variation of Foldiak's trace rule [31] to model the visual response to altered temporal exposure. We first show that this temporal association learning rule is accurate by comparing its performance to that of a similar model with hard-wired translation invariance [4]. This extends previous modeling results by Masquelier *et al.* [73] for models of V1 to higher levels in the visual recognition architecture. Next, we test the robustness of translation invariance learning on single cell and whole population responses. We show that even if single cell translation invariance is disrupted, the whole population is robust enough to maintain invariance despite a large number of mis-wirings.

The results of this study provide insight into the evolution and development of transformation invariance mechanisms in the brain. It is unclear why a translation invariance learning rule, like the one we modeled, and those confirmed by [9, 10, 33], would remain active after development. We have shown that the errors associated with a continuously active learning rule are negligible, and thus it may be simpler to leave these processes active than to develop a mechanism to turn them off.

Extending this logic to other transformations is interesting. Translation is a *generic* transformation; all objects translate in the same manner, so translation invariance, in principle, can be learned during development for all types of objects. This is not true of "non-generic" or *class-specific* transformations, such as rotation in depth, which de-

pends on the 3-D structure of an individual object or class of objects [7, 11, 79]. For example, knowledge of how 2-D images of faces rotate in depth can be used to predict how a new face will appear after a rotation. However, knowledge of how faces rotate is not useful for predicting the appearance of non-face objects after the same 3-D transformation. Many transformations are class-specific in this sense[3]. One hypothesis as to why invariance-learning mechanisms remain active in the mature visual system could be a continuing need to learn and refine invariant representations for more objects under non-generic transformations.

Disrupting rotation in depth has been studied in psychophysics experiments. Wallis and Bulthoff showed that training subjects with slowly morphing faces, disrupts viewpoint invariance after only a few instances of altered training [32, 77]. This effect occurs with a faster time course than observed in the translation invariance experiments [9]. One possible explanation for this time discrepancy is that face processing mechanisms are higher-level than those for the “greeble objects" and thus easier to disrupt. However, we conjecture that the strong, fast effect has to do with the type of transformation rather than the specific class of stimuli.

Unlike generic transformations, class-specific transformations cannot be generalized between objects with different properties. It is even possible that we learn non-generic transformations of novel objects through a memory-based architecture that requires the visual system to store each viewpoint of a novel object. Therefore, it is logical that learning rules for non-generic transformations should remain active as we are exposed to new objects throughout life.

In daily visual experience we are exposed more to translations than rotations in depth, so through visual development or evolutionary mechanisms there may be more cells dedicated to translation-invariance than rotation-invariance. We showed that the size of a population of cells has a significant effect on its robustness to altered training, see Figure 4. Thus rotation invariance may also be easier to disrupt, because there could be fewer cells involved in this process.

[3]Changes in illumination are another example of a class-specific transformation. These depend on both 3-D structure and material properties of objects [11].

Two plausible hypotheses both point to rotation (class-specific) versus translation (generic) being the key difference between the Wallis and Bulthoff and Cox *et al.* experiments. We conjecture that if an experiment controlled for variables such as the type and size of the stimulus, class-specific invariances would be easier to disrupt than generic invariances.

This study shows that despite unavoidable disruptions, models based on temporal association learning are quite robust and therefore provide a promising solution for learning invariance from natural vision. These models will also be critical in understanding the interplay between the mechanisms for developing different types of transformation invariance.

3.5 Acknowledgements

This work was supported by the following grants: NSF-0640097, NSF-0827427, NSF-0645960, DARPA-DSO, AFSOR FA8650-50-C-7262, AFSOR FA9550-09-1-0606

Chapter 4

Class-specific transformation invariance implies modularity of the ventral stream

In everyday visual experience class-specific transformations like 3D rotation severely distort the appearance of objects and yet, most of the time, they are easily discounted by the human visual system. Unlike image-based (2D affine) transformations, which can be learned from experience with any objects, invariance to these transformations can only be learned from similar objects. It follows that a wide class of popular models of the ventral stream, hierarchical view-based models, cannot employ generic circuitry to solve the task of invariantly recognizing unfamiliar objects from a single example view. Class-specific invariance can be obtained by separating the processing of objects that transform differently from one another into different domain-specific modules. Importantly, each module can perform analogous computations to all the others and could be wired-up by learning or evolution in the exact same manner (e.g. association of temporally adjacent views), but each needs to operate on a different domain in order to work. The extent to which invariance learned on some objects of a class transfers to other objects in the class predicts human performance on tests of generalization from a single example view. We propose a quantitative "niceness" index which

can be measured for any class of objects. The index predicts poor performance on 3D-rotation invariant recognition of the paperclip stimuli used in some psychophysics experiments and stronger (but not perfectly invariant) performance for faces. A range of other object classes fall in between these extremes. A strong modularity conjecture argues that the index predicts which object classes are expected to have domain-specific modules in the ventral stream and which are not. It suggests a monotonic relationship between the index and the size of the corresponding module as observed in fMRI imaging.

4.1 Introduction

The ventral stream contains discrete patches dedicated to the visual processing of specific visual categories—e.g. faces [12, 13, 80], scenes [81], words [82, 83], and bodies [84]. Despite decades of study the underlying computational reason that the brain adopts this modular architecture remains unknown; the role—if any—played by modularity in the ventral stream's recognition algorithm has not been determined. In this chapter we argue that the modular architecture is necessary. Dedicated processing modules may exist because they are the only way that the brain could accomplish some of its recognition tasks.

For the purposes of this report, a module is a localized patch of cells in the ventral stream that respond selectively to a specific class of objects[1]. The example for which the most detailed experimental data is available—and consequently our primary motivating example—is the macaque face-processing system [14]. Many of the regions found in human neuroimaging experiments with faces, bodies, and words are also modules under our definition. These include the fusiform face area (FFA), the extrastriate body area (EBA), the visual word form area (VWFA) and others. We believe that our use of the word "module" is relatively standard in the field; however to avert semantic

[1]This usage of the word module is, unfortunately, at odds with HW-module, as used in this thesis's introduction and discussion. This chapter was written before those were. When the present chapter was written we had not yet settled on a name. This paper refers to the same class of networks as considered in the introduction and discussion, but it calls them "hierarchical view-based models". It's actually a good name in this context, just not general enough for all the things we wanted to say about them elsewhere.

confusion we explicitly note that early visual areas, e.g. V1 and V2, are not modules with our definition. The lateral occipital complex (LOC) is also not a module in this sense.

The purpose of this report is to propose a new explanation at the level of computational theory (see [85]) for why there are domain-specific modules in the ventral stream. The argument is quite general, but for ease of exposition we will mostly develop it in the context of an extended example: 3D rotation-invariant face identification and the face-specific patches in macaque visual cortex. After developing the argument in that setting we explain how it generalizes to other modules.

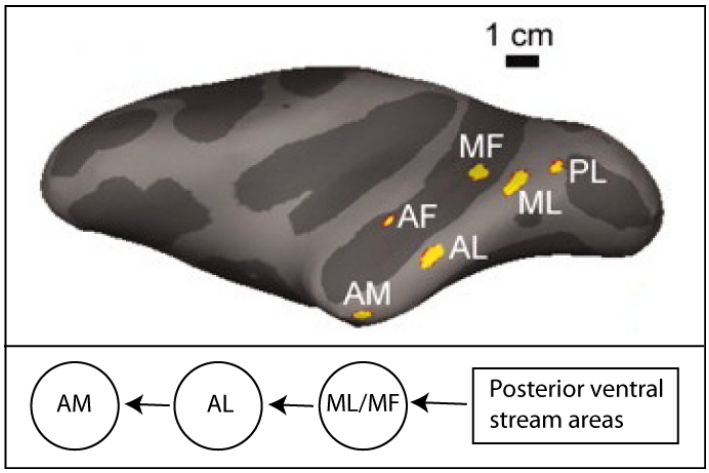

Figure 4-1: Layout of face-selective regions in macaque visual cortex, adapted from [15] with permission.

In macaques, there are 6 discrete face-selective regions in the ventral visual pathway, one posterior lateral face patch (PL), two middle face patches (lateral- ML and fundus- MF), and three anterior face patches, the anterior fundus (AF), anterior lateral (AL), and anterior medial (AM) patches [13, 14]. At least some of these patches are organized into a feedforward hierarchy. Visual stimulation evokes a change in the local field potential ~ 20 ms earlier in ML/MF than in patch AM [15]. Consistent with a hierarchical organization involving information passing from ML/MF to AM via AL, electrical stimulation of ML elicits a response in AL and stimulation in AL elicits a response in AM [86]. In addition, spatial position invariance increases from ML/MF to AL, and increases

further to AM [15] as expected for a feedforward processing hierarchy.

The firing rates of cells in ML/MF are most strongly modulated by face viewpoint. Further along the hierarchy, in patch AM, cells are highly selective for individual faces but tolerate substantial changes in viewpoint [15]. Freiwald and Tsao conjectured that a goal of this processing hierarchy is to compute a representation of faces that is invariant to 3D rotation-in-depth, and that this representation may underlie face identification behavior [15].

Our argument follows from a few simple and widely held premises:

1. The computational goal of the ventral stream is to compute representations that are at once selective for individual objects and invariant to identity-preserving transformations of their appearance.

2. The ventral stream implements a feedforward recognition algorithm. Note: This does not preclude an important computational role for recurrent connections; we only need to assume that the ventral stream can solve the "core" problems of object recognition in its feedforward mode [17, 87].

3. The ventral stream's recognition algorithm is view-based and hierarchical [41, 88, 89].

The main thrust of our argument—to be developed below—is this: The ventral stream computes object representations that are invariant to transformations. Some transformations are *generic*; the ventral stream could learn to discount these from experience with any objects. Translation and scaling are both generic (all 2D affine transformations are). However, it is also necessary to discount many transformations that do not have this property. Many common transformations are non-generic; 3D-rotation-in-depth is the primary example we consider here.

From a single example view it is not possible to achieve a perfectly invariant representation with respect to a non-generic transformation. These transformations depend on information that is not available in a single image, e.g. the object's 3D structure. Despite this, approximate invariance can still be achieved using prior knowledge of how

similar objects transform. Many non-generic transformations are *class-specific*. The transformation of object appearance caused by a 3D-rotation-in-depth is not the same 2D transformation for two objects with a different 3D structure, e.g. a face simply does not rotate like a car. However, within a restricted class where all the objects have similar 3D structure, all objects do rotate (approximately) the same way. Faces are the prototypical example of such a *nice* class, where all its members transform similarly (cf. "linear" classes in [79, 90])).

The conclusion of our argument will be the claim that any view-based recognition architecture that discounts class-specific transformations must be modular. The circuitry that underlies view-invariant unfamiliar face-identification must be separated from the circuitry that computes the same transformation invariance for other objects. We propose that the need to discount class-specific transformations is the underlying computational reason for modularity in the ventral stream.

4.2 Results

4.2.1 Hierarchical view-based models

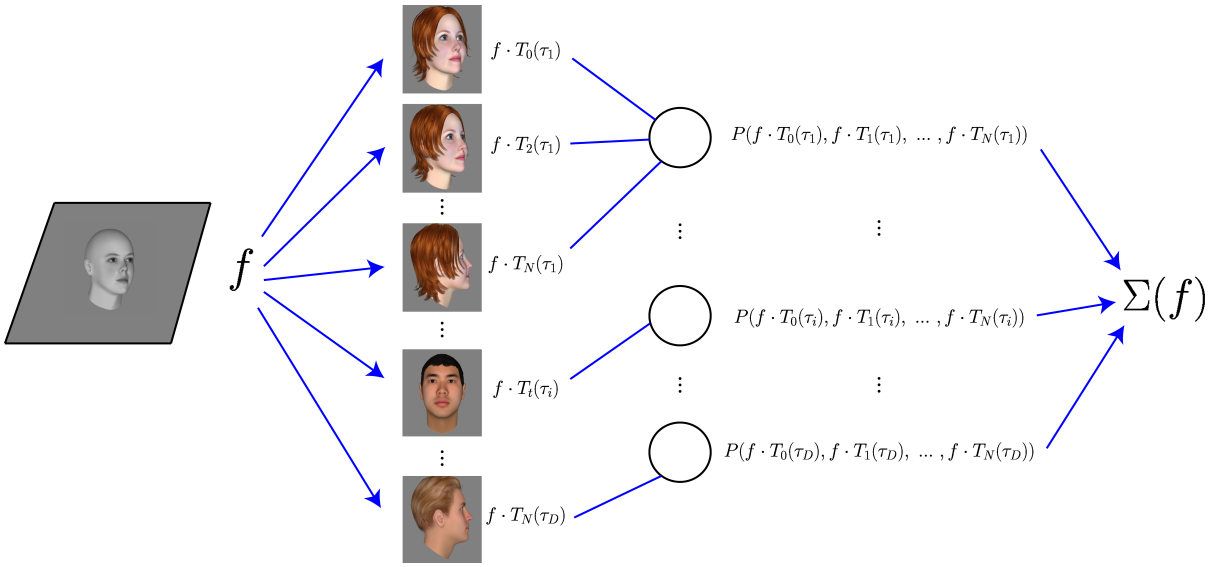

Figure 4-2: Illustration of a hierarchical view-based model.

A wide range of view-based object recognition models are compatible with our premises. View-based models encode images by their similarity to a set of stored templates—pooled over transformations. Hubel and Wiesel's model of how V1 complex cell receptive fields are built from simple cell receptive fields is an example of this sort of computation. Simple cells are (approximately) tuned to oriented bars at specific positions in the visual field. Complex cells are orientation-tuned, but less sensitive to the exact position of the stimulus. Hubel and Wiesel proposed that complex cells pool over several simple cells with the same preferred orientation but slightly different preferred positions [21]. In our transformation-focused language we can say that a complex cell pools over translations of a particular oriented template.

With the term "hierarchical view-based models" we intend to refer to the whole class of models that extend Hubel and Wiesel's model hierarchically, perhaps pooling over different transformations besides translation. This is a very wide class of models that includes Fukushima's neocognitron [27], the HMAX family of models [3, 4, 91], convolutional neural networks [28], VisNet [92], and many others (e.g. [93–96]). By necessity, the simulations in this paper use a particular model but the computational properties of generic and class-specific transformations are model-independent. None of our claims here rely on a specific model choice; all the simulation results could be replicated using any hierarchical view-based model.

View-based models work by storing templates which can be thought of as frames from the movie of an object undergoing a transformation. The *signature* that arises at the top level can be thought of as measuring the similarity of the input pattern to the templates—pooled over their previously-viewed appearances. Numerous algorithms have been proposed to solve the problem of how to wire up these models through temporal association [31, 35, 74, 75, 92, 97]. There is also psychophysical and physiological evidence that visual cortex employs a temporal association strategy [9, 10, 32, 33, 77]. None of our claims depend on the details of the learning process, thus we can restrict the subsequent analysis to the "mature" form of these models.

Consider a mechanism that stores frames as an initial pattern transforms under the action of a specific transformation (such as rotation). This is the "developmental" phase

of learning the templates. At run time, the model computes the similarity of the viewed image to each of the templates under all their stored transformations—e.g. with a normalized dot product. This set of similarities is analogous to the responses of simple cells in Hubel and Wiesel's model. In the second stage (analogous to complex cells), each element of the signature is computed by applying a pooling function over the similarities of the viewed image to all the stored transformations of one base template—e.g. computing the max or the average.

The entire view-based model computation can be repeated hierarchically by taking the signature at the previous layer for the input and tempate-encoding used in the similarity computation at the subsequent layer. The properties of generic and class-specific transformations affect single layer architectures just as they do deeper architectures. Thus, for simplicity, we use a 1-layer view-based model for our simulations. [11] contains the analogous simulation results for a more complex hierarchical model (HMAX).

4.2.2 Generalization from a single example view

We simulated tests of initial invariance for unfamiliar objects. 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 object as the reference (the target), or it may depict a distractor object. 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)—analogus to d' for the corresponding behavioral experiment [7, 58]. Figures 4-3, 4-4, and 4-5 show 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.

Variations on the modeled task are commonly used in psychophysics experiments. For example, Bülthoff et al. measured how many degrees of 3D rotation that human observers could generalize from a single example view of an unfamiliar object [41]. Logothetis et al. performed a similar experiment with two classes of unfamiliar objects and monkey subjects [88]; they also measured the 3D rotation invariance of neurons in IT cortex [25]. There are also many of the analogous experiments for other transformations in the literature, e.g. translation [43, 48] and in-plane rotation [49].

4.2.3 Invariance to generic transformations

Figure 4-3 shows the results of a test of translation-invariant novel face recognition (left column). The view-based model performs perfectly on this task (blue curves) while a control pixel-based model's performance declines rapidly with translation (red curves). The template views could be acquired either from faces (top row) or from random noise patterns (bottom row) with no change in performance. Likewise, random noise patterns could be recognized using either face templates or random noise templates (right column).

We introduce the term *generic transformation* to indicate a transformation for which invariance learned on any templates exactly transfers to all other objects. That is, a transformation is generic if it satisfies the following two conditions:

1. It generates transformed images from a single given image of an object.

2. Invariance to it can be learned for any new image from transformations of any image.

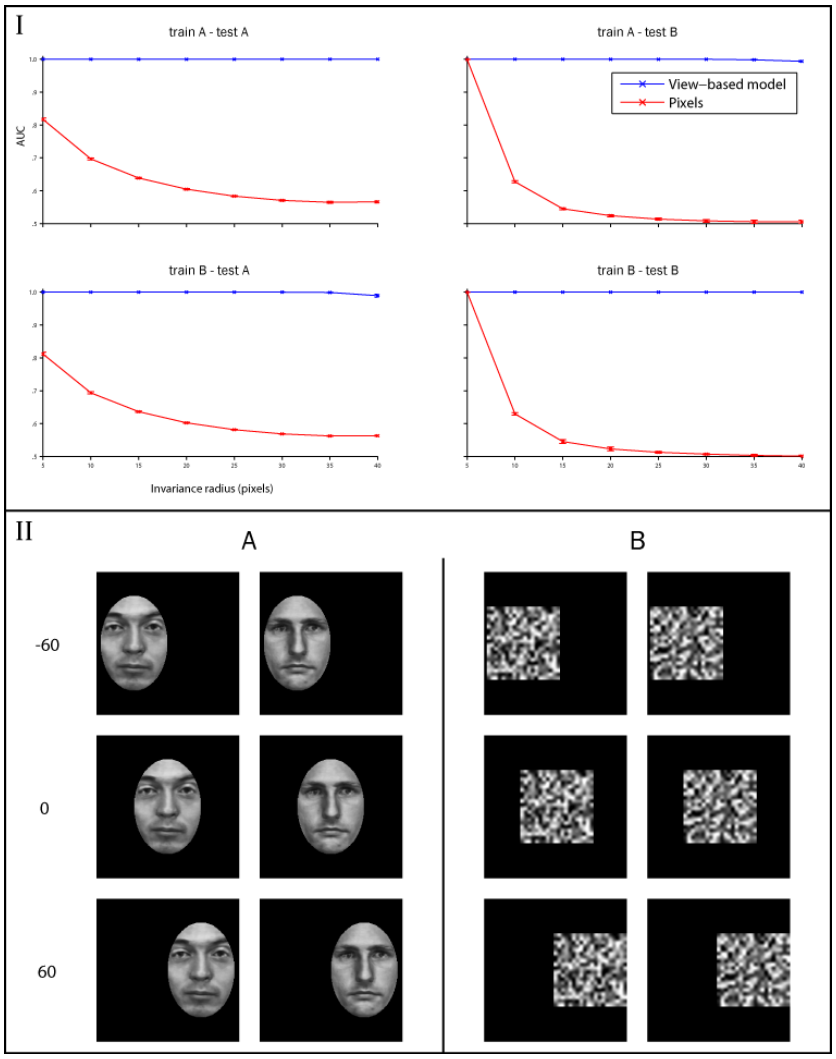

Figure 4-3: Generic invariance to translation. Bottom panel (II): Example images from the two classes. The faces were obtained from the Max-Planck Institute dataset [66] and then contrast normalized and translated over a black background. Top panel (I): The left column shows the results of a test of 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.

See supplementary section 4.6.1 for additional discussion of generic transformations. The results in figure 4-3 are an empirical demonstration that translation is a generic transformation. [36] contains a mathematical argument that all 2D affine trans-

formations are generic.

4.2.4 Approximate invariance to class-specific transformations

Viewpoint (rotation in depth)

Non-generic transformations like 3D rotation-in-depth depend on information that is not available in a single image. Perfect invariance to non-generic transformations is not possible for unfamiliar objects. However, approximate invariance can still be achieved as long as the template objects transform similarly to the test objects. One view of this is to say that the missing information in the object's 2D projection is similar between template and test objects. For example, 3D rotation is a non-generic transformation—as a map between projected 2D images it depends on the object's 3D structure. If the template and test objects have the same 3D structure then the transformation learned on the template will apply exactly to the test object. If they differ in 3D structure then the error incurred is a function of the difference between their 3D structures (see Supp. Mat. 4.6.2).

Faces all approximately share a common 3D structure; thus approximate invariance to 3D rotation of faces can be transferred from template faces to unfamiliar faces (fig. 4-4—upper left plot). This only works if the templates are actually views of faces. Using templates drawn from 3D rotations of non-face objects does not yield approximate invariance or boost discriminability beyond the level achieved by using the raw pixels for classification (fig. 4-4—left column, bottom two plots). These results are an empirical demonstration that 3D rotation is a class-specific transformation for which faces are a particularly nice class.

Other object classes besides faces are nice with respect to 3D rotation. We used 3D graphics software to render images of two additional object classes under 3D rotations (fig. 4-4—class B and C objects). Just as with faces, within each class there is a shared gross 3D structure, e.g. class B objects are all conical and have a centered protrusion near the top; class C objects all have a pyramidal protrusion in the center and two wall-like bumps on either side. Individuals were defined by smaller variations in 3D structure

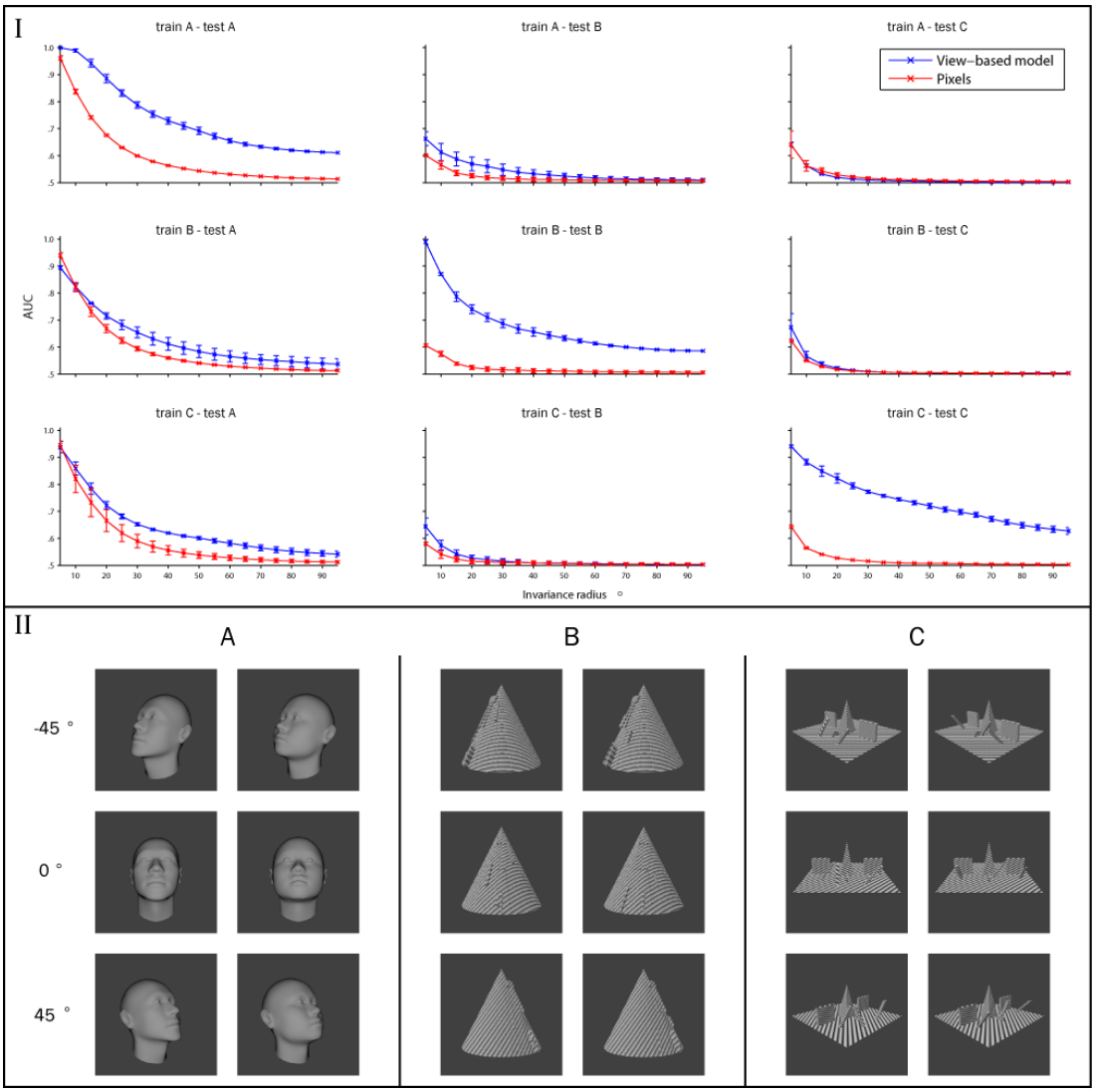

Figure 4-4: Class-specific approximate invariance to rotation in depth. Bottom panel (II): Example images from the three classes. Top panel (I): The left column shows the results of a test of 3D rotation invariance on faces (class A), the middle column shows results for class B and the right column shows 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.

within each class e.g. the position and orientation of several smaller protrusions. View-based models were approximately invariant to 3D rotation whenever they were built using templates drawn from the same object class that they were tested on (fig. 4-4—

plots on the diagonal). Performance of the view-based models trained using templates from the wrong object class followed the performance of the pixel-based model, that is, they declined to chance level (AUC = .5) by $\sim 25^\circ$ of rotation. Thus object classes B and C are also nice with respect to 3D rotation. This simulation also demonstrates that the combined class consisting of all three object classes together is not nice. Encoding unfamiliar members of that class relative to template views of its other members will only sometimes (1/3 of the time) yield an approximately invariant signature.

Illumination

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 4-5 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. 4-5—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. 4-5—off diagonal plots). This simulation demonstrates that these object classes are nice with respect to illumination transformations. However, the weak performance of the view-based model on the silver objects indicates that it is not as nice as the others. 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.

Body pose

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 $t_p : B \rightarrow \mathbb{R}^d$.

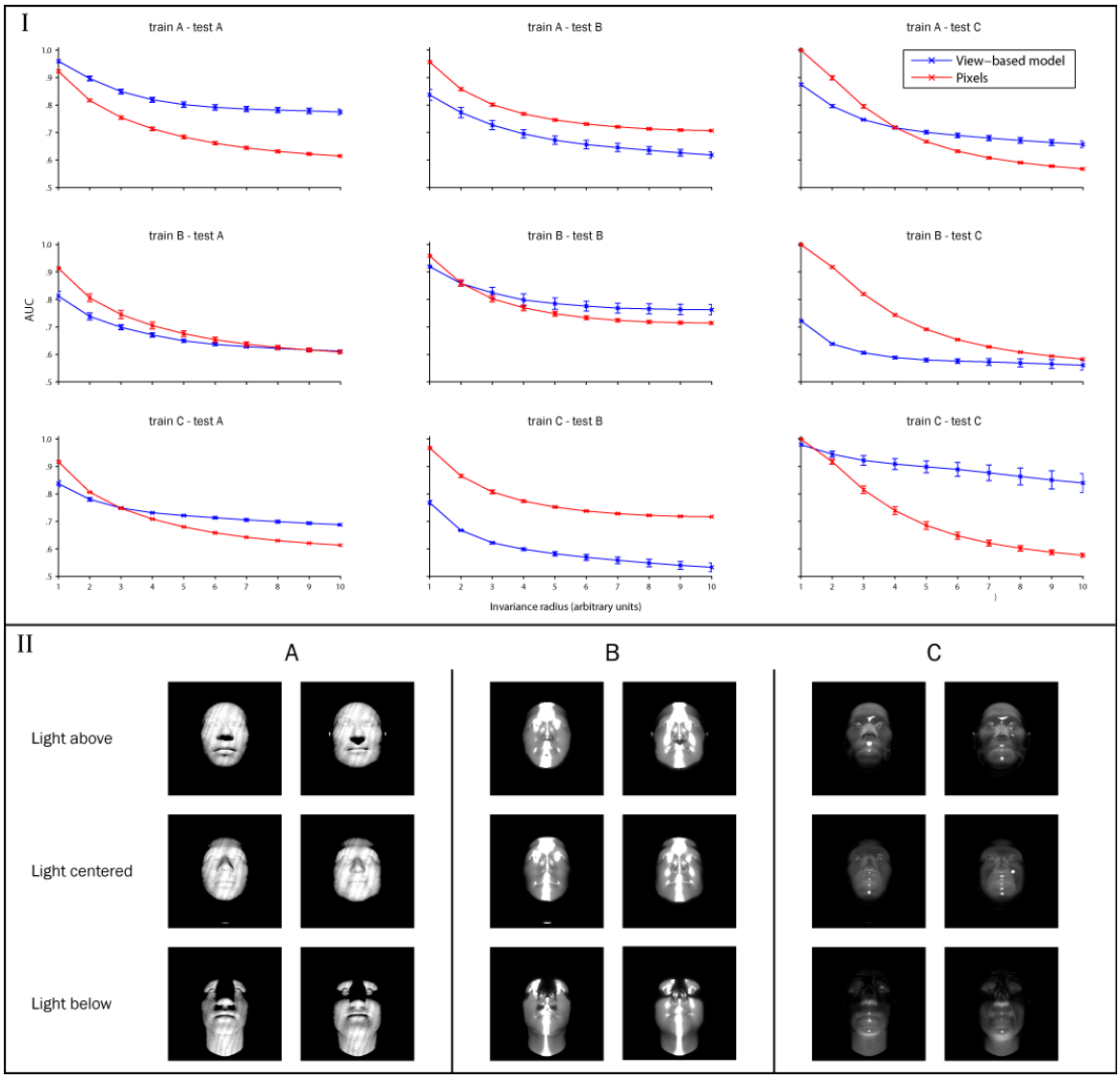

Figure 4-5: Class-specific approximate invariance to illumination changes due to movement of the light source. The light source moved vertically in front of the object. Same organization as figure 4-4.

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

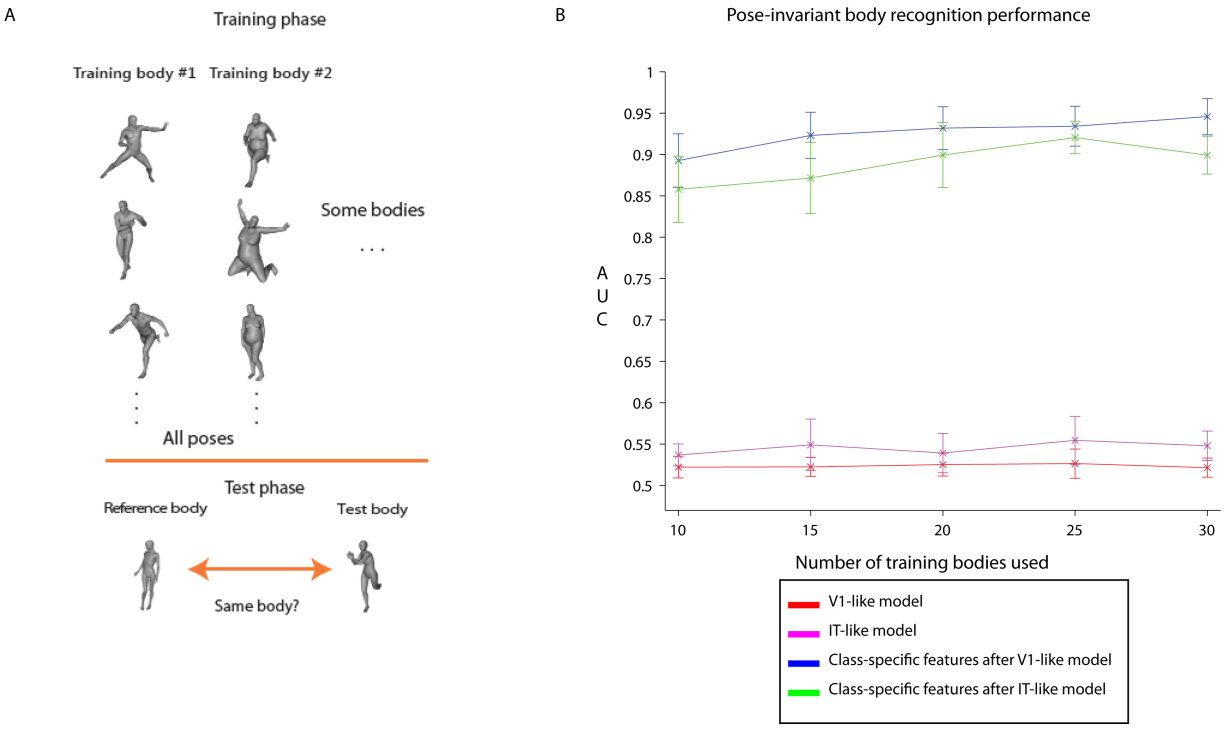

Figure 4-6: **A.** Example images for the pose-invariant body-recognition task. The images appearing in the training phase were used as templates. The test measures the model's performance on a same-different task in which a reference image is compared to a query image. 'Same' responses are marked correct 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.

$$\mu(x) = \begin{pmatrix} \max(\langle I, t_1(\tau_1)\rangle, \langle I, t_2(\tau_1)\rangle, \ldots, \langle I, t_n(\tau_1)\rangle) \\ \max(\langle I, t_1(\tau_2)\rangle, \langle I, t_2(\tau_2)\rangle, \ldots, \langle I, t_n(\tau_2)\rangle) \\ \vdots \\ \max(\langle I, t_1(\tau_m)\rangle, \langle I, t_2(\tau_m)\rangle, \ldots, \langle I, t_n(\tau_m)\rangle) \end{pmatrix} \tag{4.1}$$

For this experiment we used a Gaussian radial basis function to model the S-unit response.

$$\langle I, t_i(\tau_j)\rangle = \exp\{\sigma * \sum((I - t_i(\tau_j))^2)\} \tag{4.2}$$

Where σ is the Gaussian's variance parameter.

The model can be made hierarchical. It 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 C2b)[52]—referred to in fig. 4-6 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. 4-6).

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. 4-6).

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 cf. commentaries on [98] in the same journal issue). 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.

Discussion

Fundamentally, the 3D rotation of an object class with one 3D structure, e.g. faces, is not the same as the 3D rotation of another class of objects with a different 3D structure. Generic circuitry cannot take into account both transformations at once. The templates used by a view-based model to recognize unfamiliar faces despite 3D rotation must also be faces. We conjecture that the reason the macaque face-processing system is segregated into discrete patches is to facilitate class-specific pooling. To ob-

tain a rotation-invariant representation in an anterior face patch (e.g. patch AM) it is necessary to ensure that all the AM cells' inputs transform like faces (here we mean necessary, and sufficient, on the computational level of analysis, in the sense that we have an "if and only if" theorem. See [1]). This connectivity is most simply achieved if the upstream face-selective cells are nearby one another in a particular patch of cortex.

We call this conclusion the *weak modularity implication*. If the premises are satisfied, i.e. the ventral stream's core view-based feedforward operation computes as-invariant-as-possible representations for novel objects, then it must separate the processing of some object classes from others in order to achieve class-specific generalization for unfamiliar objects. The following sections expand on this argument and articulate a *strong modularity conjecture*.

4.2.5 Nice classes

Niceness is a graded notion. It is possible for one set of objects and a transformation to be more or less nice than another. For example, we saw in figure 4-4 that the set of faces was nicer with respect to 3D rotation than the combined set of class A, B and C objects. In figure 4-3, we saw that faces were even nicer with respect to translation. We define a quantitative index of niceness that can be computed directly from the sequence of images depicting a set of objects transforming. Since generic transformations apply the same way to all objects we can say that the maximum possible niceness is achieved for generic transformations. Intuitively, the set consisting of any single object with respect to any transformation would also be maximally nice (though that case is not accomodated in our simulations since we always require the template and test sets to be independent).

We have demonstrated that approximate invariance to some non-generic transformations can be obtained by processing each nice class separately (figures 4-4 and 4-5). Consider an object recognition system with a number of distinct modules. One extreme example is the "grandmother cell" system that has a separate module for every object. This system would maximize the total niceness over all its modules. However,

it would defeat the whole purpose of having a view-based model; if every object is in a separate module from every other object then there could never be any transfer of invariance from templates to unfamiliar objects. At the other extreme we have the "fully distributed" system with only one, highly not-nice, module that processes all objects. Most of the elements of its signature would indicate similarities to templates that do not transform similarly to the test object. This signature would change drastically when the unfamilar object transforms.

The claim that the ventral stream falls somewhere on the spectrum between the two extremes of grandmother cells and completely distributed codes is not controversial. The argument we have developed here makes some stronger predictions than this though. In some cases we can even make specific statements about which modules are expected and which are not.

The "paperclip" objects used in many psychophysics and physiology experiments on 3D rotation-invariant recognition (e.g. [25, 41, 88]) are defined entirely by their 3D structure. They do not have any relevant 3D structure in common with one another. Thus the class of paperclip objects is very not-nice with respect to 3D rotation (see the table of niceness values below). This is why experience with invariantly identifying template paperclips does not facilitate subsequent invariant identification of unfamiliar paperclips [41, 88].

4.2.6 The strong modularity conjecture

Object representations may be assigned to modules in order to maximize niceness. Consider an object recognition system with a number of modules. When a new object is learned, add its representation (set of associated template images) to the module for which its addition gives the highest niceness. Most of the time, adding a new object representation to a module will decrease its niceness. If there is no module to which the object can be added without decreasing niceness below a threshold then create a new module to contain that object's representation. After repeating this procedure for many objects—sampled according to the distribution of objects encountered in natural

vision—there will be a few large modules with many objects in them and many small modules with very few, or just one, object.

The largest module generated by this procedure would almost certainly be a face module. Modules representing human bodies—perhaps invariantly to posture—would also be one of the larger modules. If the objects are sampled according to their distribution in natural vision, then orthography in a literate viewer's native language would also likely have a large module. The orthography module might be invariant to transformations induced by differences in viewing angle or even more exotic—but common—transformations like font or handwriting styles. The smallest modules would contain objects that do not transform similarly to any other objects. The paperclip objects commonly used for experiments would end up being isolated in their own modules (they are not a nice class with respect to 3D rotation).

Neuroimaging methods like fMRI have limited resolution, thus only the largest modules will be visible to them. The strong modularity conjecture predicts that the region of cortex containing the smaller modules will appear in neuroimaging experiments not to be specifically tuned to any particular object class—and it will have the additional property that it will not respond strongly to objects represented by the large modules. That is, the predictions of the strong modularity conjecture are in accord with the results of neuroimaging experiments that show dedicated face regions—e.g. FFA, OFA [12, 99]—body regions—EBA, FBA [84, 100]—and an orthography region—VWFA [82]—as well as experiments that report strong responses to other object stimuli in LOC [101].

The strong modularity conjecture predicts that additional modules will be found with higher resolution scans. In accord with this prediction, a recent high-field (higher resolution) primate fMRI study reported that there is a small fruit-selective patch in visual cortex [80]. Another implication of our argument is that the paperclip-tuned cells described in [25] would not have been localized in a discrete patch of cortex. To our knowledge, the spatial distribution of these cells (or similar ones) was never measured; so this may be a true experimental prediction. In general, the strong modularity conjecture predicts a monotonic relationship between an object class's niceness and the size of its corresponding module.

4.3 Discussion

We have argued that the ventral stream implements a modularity of *content* rather than *process*. The computations performed in each module can remain quite similar to the computations performed in other regions. Indeed, the connectivity within each region can be wired up in the same way, e.g. through temporal association. The only difference across areas is the object class (and the transformations) being encoded. In this view, the ventral stream must be modular in order to succeed in the tasks with which it is faced.

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 and Heejung Kim for assistance with the body-pose simulations. 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 research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation.

4.4 Methods

4.4.1 Stimuli

All objects were rendered with perspective projection.

Generic fig: there were 100 faces and 100 random noise patterns randomly partitioned into 30 for templates and 30 for testing.

3D rot: there were 40 faces, 20 class B, 20 class C. Randomly picked 10 templates and 10 test for each of 20 cval runs.

Ilum: 40 wood, 40 silver, 40 glass. Randomly picked 20 for templates and 20 for test, in 20 cval runs.

3D rotation: We obtained 3D models of objects from 3 different classes and used Blender [102] to render images of them at each orientation in 5° increments from -95° to 95°. Class A consisted of faces; we used Facegen [103] to create the 3D models. The faces were all untextured so the task required the use of shape to identify individuals. Class B and C objects were modeled with Blender.

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/).

The 3D meshes and material data for all the faces and other objects created for our simulations will be made freely available from http://cbcl.mit.edu (or as supplementary data).

4.4.2 Niceness index

We define a compatibility function $\psi(A, B)$ to quantify how similarly two objects transform, where A_i is i_{th} frame/image of object A. $i = 1 \ldots m$ and B_i is i_{th} frame/image of object B. $i = 1 \ldots m$.

Then we can define the *niceness index* to be the average $\overline{\psi}$ over all pairs in the class.

4.5 Table of niceness values for various classes

Object class	Transformation	$\overline{\psi}$
fig. 4-4-class A (faces)	Rotation in depth	0.57600
fig. 4-4-class B	Rotation in depth	0.95310
fig. 4-4-class C	Rotation in depth	0.83800
All fig. 4-4	Rotation in depth	0.26520
Chairs (Digimation)	Rotation in depth	0.00540
Wire objects [88]	Rotation in depth	-0.00007
COIL-100 [104]	Rotation in depth	0.00630
Glass statues-fig. 4-5	illumination	0.56320
Sliver statues-fig. 4-5	illumination	0.35530
Wood statues-fig. 4-5	illumination	0.53990

4.6 Supplementary material

4.6.1 Generic transformations

For X the set of all 2D images and $T : X \to X$ an invertible transformation, we say that T is generic if for any pair of images $x_1, x_2 \in X$

$$< T(x_1), x_2 >=< x_1, T^{-1}(x_2) > \tag{4.3}$$

Where $< \cdot, \cdot >$ indicates the template response function. For example, $< x, \tau >$ could model the response evoked by x for a simple cell tuned to τ. Different models use different template response functions; the simulations in the present paper use the normalized dot product.

The signature obtained by pooling over template images depicting a generic transformation of any familiar object is invariant, for example, $\max_t(< x, T_t(\tau) >)$ is invariant. Any other pooling function besides the $\max$ could also be used. The $T_t(\tau)$ could be

frames of the video of a transforming object. Or the template responses could all be computed by a convolution (as in [28]). The specific templates and their relationship to the test image does not enter the argument; in principle, cells that are particularly selective for the test object are not required to obtain a signature that is invariant to generic transformations.

4.6.2 Approximate invariance to 3D-rotation

Hierchical view-based models can achieve approximate invariance for non-generic transformations as long as the template objects transform similarly to the test objects. One view of this is to say that the missing information in the object's 2D projection is similar between template and test objects. For example, 3D rotation—as a map between projected 2D images depends on the object's 3D structure. If the template and test objects have the same 3D structure then the transformation learned on the template will apply exactly to the test object. If they differ in 3D structure then the error incurred depends on the difference between their 3D structures.

See [1] for this argument. A version of it will also be included in the supplementary information of the journal version of this chapter.

Chapter 5

View-invariance and mirror-symmetric tuning in the macaque face-processing network

Joel Z. Leibo, Fabio Anselmi, Andrea Tacchetti, Jim Mutch, Winrich A. Freiwald, and Tomaso Poggio

The ventral stream rapidly computes image representations which are simultaneously tolerant of identity-preserving transformations and discriminative enough to support robust recognition and memory-indexing. One algorithmic approach to this task involves subjecting the image to a cascade of selectivity-increasing AND-like operations (filtering) and tolerance-increasing OR-like operations (pooling) [1, 3, 4, 21, 27, 28, 94]. Many networks of this type have been used as models of the ventral stream [11, 52, 60]. The macaque face processing system, with its progression from view-specific neurons in the middle lateral and middle fundus (ML/MF) patches to view-tolerance in the anterior medial (AM) patch [15], seems like a good candidate for modeling along these lines. However, such an explanation renders mysterious the finding of mirror-symmetric orientation tuning curves in the anterior lateral (AL) face patch: the intermediate step between view-specificity and view-tolerance. Here we show that a

new biologically-plausible unsupervised learning rule for models in this class—related to Oja's [105] and Foldiak's [31] rules—yields a network which computes a representation resembling AL in its penultimate layer, prior to computing a view-tolerant representation in its AM-like final layer. Furthermore, we show that it leads to the implication that neuronal tuning properties at all levels of the ventral stream hierarchy are given by solutions to a particular eigenvalue equation we named the *cortical equation*. It can also be solved with parameters modeling other parts of the ventral stream besides the face patches. In particular, we show that in the case of small receptive fields, as in primary visual cortex, its solutions resemble Gabor wavelets.

5.1 View-based HW-modules

Consider the special case of a filtering and pooling circuit proposed by Hubel and Wiesel as a model of V1 simple and complex cell connectivity [21]. For I, the input image, t^k a stored template, $\{g_{\vec{x}}\}$ a family of translations parameterized by i, and σ a nonlinearity (e.g., a sigmoid), the k-th element of the *signature* vector μ is given by

$$\mu^k(I) := \sum_i \sigma(I \cdot g_{\vec{x}_i} t^k). \tag{5.1}$$

This way of writing the basic *HW-module*, named for Hubel and Wiesel's contribution, highlights that the pooling is done over a particular family of transformations—translations in this case. It can be generalized by considering other transformations of the stored templates. It can be shown that, under very general conditions, the output of an HW-module which pools over any locally compact group G (e.g. affine or translation, scaling, in-plane rotation), will be invariant to that transformation [1]. In the case of g an affine transformations $\mu(gI) = \mu(I)$.

In order for an HW-module to be invariant, the orbit $O_k = \{gt^k, g \in G\}$ under the transformation must be stored in the preferred features of the (analogs of) simple cells—called *S-units* here. One way to obtain transformation sequences, i.e., subsets

of orbits, is by memorizing the frames of the video of a transforming object. Let V be the matrix with each frame of the video as a column v_i. For simplicity, think of a single object transforming on an white background. Thus, $v_i = g_{\vec{x}_i} v_0,\ v_0 = t$. An HW-module built in this way, using a video of a transforming template object, will be invariant to that transformation [1]—see S.I. section 2. Note that when the transformation is affine, the HW-module will be invariant, regardless of any resemblance (or lack thereof) between the stored templates and the test image. When the transformation is not affine, the HW-module cannot be guaranteed to be invariant, but its output will still tolerate stimulus transformations to some degree proportional to the similarity between the transformations of the template and test objects [1, 11, 56] (see S.I. section 2 and 3).

Now consider the macaque face-processing system. A *view-based* HW-module with C-units modeling patch AM neurons could pool view-tuned S-units modeling AL neurons. Proposals along these lines have been the subject of many previous studies (e.g., [11]), and it is known that they can yield representations with similar view-tolerance properties to patch AM. However, they cannot explain why the macaque system goes through a mirror-symmetric intermediate representation (patch AL) along the way to computing a view-tolerant representation in AM.

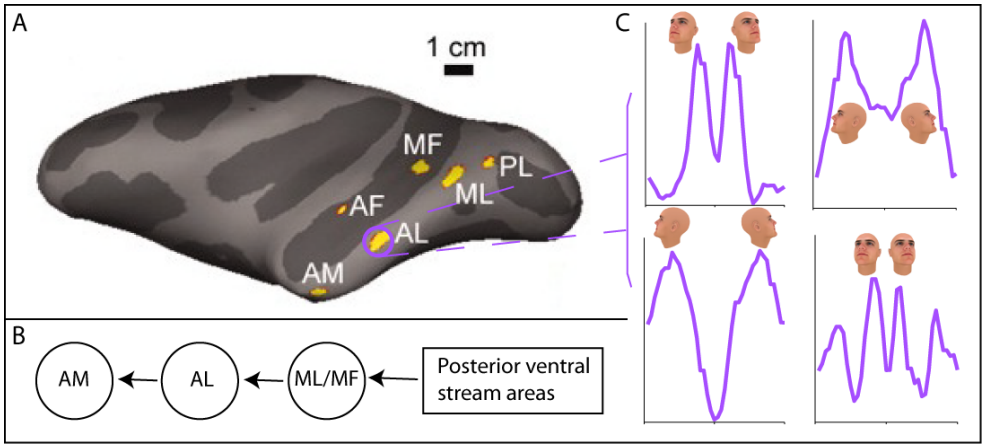

Figure 5-1: ***A.*** *Layout of face-selective regions in macaque visual cortex [15].* ***B.*** *Possible feedforward organization of the face-selective regions proposed in [15].* ***C.*** *Tuning curves of four representative AL cells. Normalized response above baseline is plotted as a function of stimulus orientation, averaged over rotations around the other two axes in 3D (data from [15]).*

In the following, we propose a new biologically-plausible unsupervised learning procedure through which HW-modules modeling the macaque face system could be learned from experience. View-based HW-modules can also be seen as *memory-based*; they store (partial) orbits. From that point of view, the HW-modules learned via our proposal can be seen as compressed versions of the standard view/memory-based HW-modules. Not only does this compression dramatically reduce storage requirements (something that was always viewed as an issue with the standard proposals), it turns out that having a mirror-symmetric representation in the layer before the view-tolerant layer can be explained as a side-effect of the compression.

5.2 Learning HW-modules

How could HW-modules develop? In the affine case, their development could be innately specified—i.e., evolution could be their learning algorithm. However, there are many reasons to favor a unified account where the circuitry for both affine and non-affine cases develop through similar mechanisms and there are only quantitative differences between them. Unified accounts are supported by the similarity of cortical structure aacross many different regions [106] and its plasticity [107]. The properties of HW-modules imply that they must be specifically tuned to templates which resemble the test objects if they are to be recognized invariantly to non-affine transformations, e.g., view-based HW-modules [11]. Despite this difference between the affine regime, where *generic* circuitry suffices, and the non-affine regime, where circuitry must be *class-specific*—as we show below, there is a single unified learning procedure which can be used to predict properties of both.

We propose the following biologically-plausible unsupervised learning procedure—first stating it succinctly, then explaining its motivation. Consider the continuous stream of visual experience as a sequence of frames $v_0, v_1, \ldots, v_i, \ldots$. Let δ be a spatiotemporal scale parameter that controls the aperture size of a simple cell both in space and time. Let V_δ^k be the matrix of frames from the video corresponding to the orbit of t^k.

The response of a learned HW-module is given by

$$\mu^k(I) := \sum_i \sigma(I \cdot \omega_i^k). \tag{5.2}$$

where the learned templates ω_i^k are the solutions to the following eigenvector-eigenvalue equation

$$V_\delta^k V_\delta^{k\mathsf{T}} \omega_i^k = \lambda_i^k \omega_i^k \tag{5.3}$$

We call (5.3), the *cortical equation*.

This procedure operates on a local spatiotemporal neighborhood, it falls into the class of unsupervised temporal-adjacency-based learning algorithms. Since adjacent frames of visual experience tend to depict the same object, these mechanisms can be used to associate the appearances of an object under different transformation conditions. The phenomenon of unsupervised temporal-adjacency learning was demonstrated in both physiology [10, 33, 108] and psychophysics [9, 32] experiments. It has been formalized in several different ways [31, 35, 97, 109], the most famous being Foldiak's trace rule.

A key parameter of any temporal-adjacency based learning rule is the temporal scale on which it operates. In our case, HW-modules also have a spatial scale so we can treat them together and consider a unified *spatiotemporal scale*. There need only be one parameter to describe the aperture size in both space and time since meaningful transformations are only observed on small spatial windows over short periods of time or on large windows over a long time. It is not useful to be invariant to fast transformations over large spatial windows or slow ones in a small visual field, these transformation are akin to noise for object recognition purposes—see supplementary information section 5.3.4. When the spatiotemporal scale is short, the HW-modules obtained will be appropriate for the generic (affine) regime, see SI section 3. That is, they will tend to learn to tolerate localized (2D) affine transformations. When the spatiotemporal scale is wide, HW-modules for the class-specific regime will arise (see figure 5-4).

Another standard assumption, which this theory also invokes, is experience evokes

synaptic plasticity in accord with Hebb's principle [110] that "neurons that fire together wire together", this concept can be written as: $\Delta w^k = \gamma (w^k)^T v_\delta$, where v_δ is a transformation sequence i.e. a column of V_δ^k, and w^k is the vector of presynaptic weights. This dynamical system converges if the weights are normalized and there is substantial experimental evidence in support of normalization in the cortex [111]. Mathematically, this implies a modification to Hebb's rule the simplest version of which has been proposed by [105]. It has been shown that Oja's version of Hebb's rule converges to the eigenvector associated to the largest eigenvalue of the covariance of its inputs, that is, the vector of presynaptic weights converges to a solution of the cortical equation.

Recall that a view-based HW-module is memory-based since it stores frames of the video depicting a face's rotation. Now consider the HW-module learned with a wide spatiotemporal scale from watching the same video. It is a compressed version of the view-based HW-module. The specific compression algorithm is principal components analysis (PCA). Figure 5-2 shows that very few components are needed to obtain the same performance as the view-based model on a difficult test of view-invariant face recognition. In principle, this compression could lead to a dramatic easing of storage requirements. It also motivates a study of the properties of the first few eigenvectors of the covariance of the video of a rotating face, i.e., solutions of the cortical equation for the case of a high-level face-specific region of the ventral stream.

Consider a developing HW-module, trained on the video of a rotating face. Since faces are approximately bilaterally-symmetric, the video will generally contain the reflection of each of its frames over the vertical midline. It turns out, in the case where this holds exactly, the solutions to the cortical equation must be either symmetric or antisymmetric (i.e., the eigenfunctions of the associated covariance operator are either even or odd); see S.I. section 5.3.6 for a proof.

An S-unit tuned to a symmetric feature will always have a symmetric orientation tuning curve. Furthermore, if the nonlinearity σ is an even function, as in the energy model [22], then S-units tuned to antisymmetric features will also always have symmetric orientation-tuning curves. Figure 5-3 shows example tuning curves from the same simulation as in figure 5-2. That is, it plots $\sigma(\omega_i^k \cdot I_\theta)$ as a function of the test face's

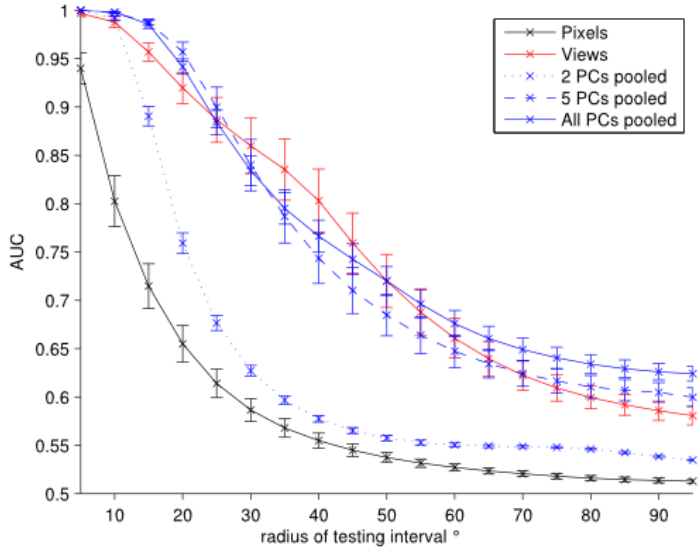

Figure 5-2: ***A.*** *Results from a test of viewpoint-invariant face identification. This test models the psychophysics task used in [66] (and many others). Test faces were presented on a gray background. The task was to correctly categorize images by whether or not they depict the same person shown in a reference image—despite changes in viewpoint. This is a test of generalization from a single example view. The abscissa shows the maximum invariance range (maximum deviation from the frontal view in either direction) over which targets and distractors were presented. The ordinate shows the area under the ROC curve (AUC) obtained for the task of recognizing an individual despite changes in viewpoint (nearest neighbor classifier). The model was never tested with any of the images that were used to produce eigenvectors (or stored frames). We averaged the AUC obtained from experiments on the same model using all 20 different reference images and repeated the entire simulation (including the developmental phase, using a different choice of 20 training faces) 5 times with different training/test splits (for cross validation). The error bars shown on this figure are 1 standard deviation, over cross validation splits.*

orientation, at angle θ, for 5 example units tuned to features with different corresponding eigenvalues. All of these tuning curves are symmetric about 0° corresponding to a frontal face.

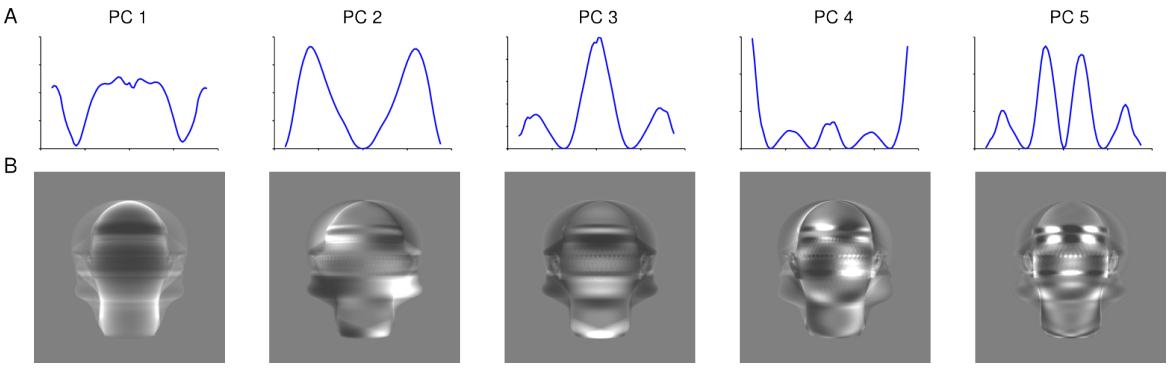

Figure 5-3: ***A.*** *The S-response $y = x^{\theta} \cdot \omega$ as a function of the test image's orientation (rotation in depth). The leftmost column is the projection onto the first eigenvector. The second column is the projection onto the second eigenvector, and so on. This simulation was done directly on the raw pixel representations with no preprocessing.* **B.** Example eigenvectors (contrast normalized) from the pixel-based (no preprocessing) simulation. The corresponding eigenvalues decrease from left to right.

Since another goal of this work was to find a unified learning procedure that could be used to wire up HW-modules serving as models of any layer of the ventral stream (not just the face patches), we studied solutions to the cortical equation at a variety of spatiotemporal scales (figure 5-4). No matter what transformation actually happens in the world, when viewed through a small enough spatiotemporal window it will always look like a translation. For example, rotation in depth of a face, viewed through a small spatiotemporal window looks like the translation of an eye (figure 5-4-D). This follows from the fact that a very wide class of transformations can be approximated by locally-affine transformations. An implication is that the developing ventral stream only needs to set up a hierarchy of layers which learn at different spatiotemporal scales—with small scales toward the beginning and wider scales later on. As a result of the statistics of natural vision, and the inherent spatiotemporal scales of different transformations, a hierarchy developing along these lines will come to learn HW-modules for the generic regime in its early layers, and HW-modules for the class-specific regime in later layers. Thus, properties of neurons in all layers of the ventral stream could be predicted by

solving the cortical equation with different parameters, and in the second regime, different template videos. Figure 5-4 shows first principal components (solutions to the cortical equation with highest eigenvalue) for two different videos of rotating heads for four different values of the spatiotemporal scale aprameter δ. In the generic regime, small δ it leads to the prediction of Gabor-like tuning, and in the case of a face-specific (rotation in depth) module, leads to the explanation of mirror symmetric tuning curves.

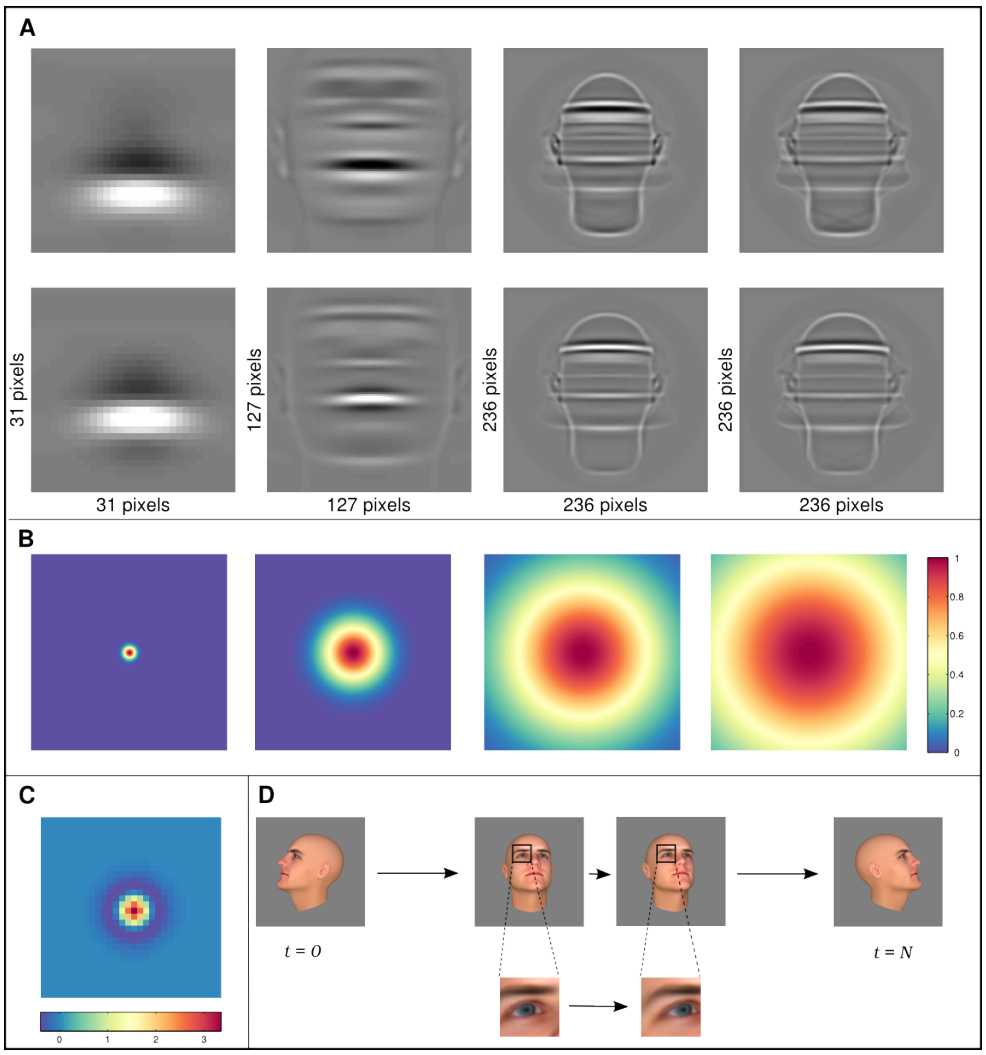

Figure 5-4: *Numerical solutions of the cortical equation with different (Gaussian) spatiotemporal scales. The spatiotemporal parameter δ was implemented by feeding the input transformation video through a Gaussian mask with equal variance in space and time (see Methods section for more details). In the figure δ increases in size from left to right.* ***A.*** *Each eigenvector of the matrix V_δ is a solution of the cortical equation. Here only the principal component associated to the largest eigenvalue is shown for the values of the aperture shown directly below it in* ***B****. The depicted Gaussian is an XY slice at the time corresponding to angle = 0, (where the maximum is achieved).* ***C.*** *The difference-of-Gaussians filter which was convolved with all frames of the video before applying the spatiotemporal aperture.* ***D.*** *Restricted to a small spatiotemporal window, the transformation induced by a 3D rotation in depth may appear as a translation.*

5.3 Supplementary Material

5.3.1 Methods: Stimuli

Faces, objects, and bodies (FOB) set: The simulation set was modified from the FOB set used in [15] by replacing all the photographs of human faces with computer-generated faces from FaceGen (Singular Inversions Inc.), rendered with Blender (The Blender foundation). This ensured that the faces in the set were more similar in appearance (no hair, etc) to the faces used for training the model. The other images in the FOB set were the same as those used in [15], except, in this case they were on a uniform gray background.

Face views (FV) set: The simulation's FV set consisted of 25 FaceGen faces rendered at the same orientations used in [15] ($0°$, $-90°$, $-45°$, $90°$, $45°$, $20°$-pitch, $-20°$-pitch, $180°$). The original FV set in [15] consisted of photographs.

The fine-angle sets: The images used for figures 5-3, 5-2, and **??** were a set of FaceGen faces rendered every $5°$ (where $0°$ was a frontal face). For all simulations, training (PCA) was done with an independent set of FaceGen faces, rendered with the same procedure. No faces used to build the model appeared in any of the test sets.

5.3.2 Methods: The test of viewpoint-invariant generalization 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 a specific 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 [7, 58]. Figure 4 (main text) 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.

5.3.3 Some background on face patches

A high proportion of cells in the macaque face patches [14] respond selectively to faces; viewed-as HW-modules, they may implement the computation of invariant representations for face-specific transformations. 3D rotation in depth is one example of a face-specific transformation [11]. In line with the property that HW-modules for non-affine transformations are only (approximately) invariant when their templates resemble (and transform similarly to) the test images. In accord with the theory, Freiwald and Tsao's experiments in a subnetwork of the macaque face processing system suggest that at least some of the face patches are organized into a feedforward hierarchy which computes a representation which is invariant to 3D rotation in depth [15].

The macaque face patches differ qualitatively in how they represent identity across head orientations [15]. Neurons in the middle lateral (ML) and middle fundus (MF) patches are view-specific, while neurons in the most anterior ventral stream face patch, the anterior medial patch (AM), are view-invariant. Puzzlingly, neurons in an intermediate area, the anterior lateral patch (AL), have mirror-symmetric orientation tuning curves. That is, neurons in patch AL typically have bimodal (or multimodal) tuning curves. For example, one might respond maximally, and roughly equally, to a left pro-

file and a right profile while responding more weakly to other orientations (fig. 5-1-C).

5.3.4 The spatiotemporal aperture

Simple cells in HW-modules are defined by their tuning functions (called templates in the text) and by their spatiotemporal aperture. We assumed throughout this chapter that the spatiotemporal aperture is hard coded (unlike the tuning functions which are learned according to the cortical equation), that is, a simple cell's position and aperture in the visual field is not acquired during development, it is a property of the cell itself.

In the simulation shown in figure 5-4, we implemented the spatiotemporal parameter by feeding the input video (i.e. the columns of V_δ in the cortical equation) through a Gaussian mask with equal variance in space and time. The effect of this is to dim the intensity in the first and last few frames of the video. The role of the aperture mask is similar in space: the images intensity is stronger in the center of the receptive field and weaker at the edges. Both in space and time the transition from low to high intensity due to the mask follows a Gaussian whose variance is the same in space and time.

As stated in the main text, we are only interested in transformations that occur quickly and over small windows, or slowly and over large windows. Rapid transformations over large windows as well as slow ones over small areas of the visual field are akin to noise for object recognition purposes. This is because a long video sequence seen through a small aperture window is likely to contain more than one object undergoing a simple transformation. HW-modules pooling over frames from this video would be "too invariant" i.e. not discriminative between different objects. On the other hand, a stream of frames that only spans a few moments but is seen through a large window likely contains only a small chuck of a transformation. HW-modules learned from such videos would not be invariant enough. This reflects a fundamental principle, related to the temporal adjacency assumption, and (probably) to the Markov nature of physics: small windows of space are self-correlated over short timescales while larger windows are self-correlated over longer timescales.

5.3.5 Hebb/Oja plasticity

We assume that experience evokes synaptic plasticity in accord with Hebb's principle [110]—the specific form of which we consider here is Oja's rule [105]. Since the naive version of Hebb's rule is unstable, it must be replaced by a suitable alternative. Oja's rule fixes the Hebb rule's instability by assuming a quadratic leakage term. Alternatively, Oja's rule can be motivated by noting that normalization mechanisms which could prevent instability of Hebbian plasticity are widespread in cortex [111], and Oja's rule can be derived as the first order expansion of the normalized Hebb rule.

Oja's rule updates a synaptic weight vector w (with convergence rate η) by

$$\Delta w = \eta(xy - y^2 w) = \eta(xx^\intercal w - (w^\intercal x x^\intercal w)w). \tag{5.4}$$

where x is the input vector, in our case the set of all transformations of an image. The original paper of Oja showed that the weights of a neuron updated according to this rule will converge to a principal component of the neuron's past inputs, equivalently: an eigenvector of the input's covariance. Informally, this can be seen by averaging the right side of (5.4), i.e. replacing $xx^\intercal$ with its covariance $\overline{C}$, and setting $\Delta w = 0$. Notice that the quadratic factor $w\overline{C}w$ is a scalar. Thus, after renaming the scalars, the synaptic weights at convergence are seen to be solutions to the eigenvector-eigenvalue equation $\overline{C}w = \lambda w$. Oja additionally showed that the basic version of his rule only gives the eigenvector corresponding to the top eigenvalue. However, it was later shown that 1. various plausible modifications of the rule can give other eigenvectors [112, 113], and 2. the basic Oja rule gives other eigenvectors beyond the first as long as the eigenvalues are relatively close to one another and there is noise [36].

5.3.6 Properties of the spectrum of the covariance of templates and their reflections

For an image $I \in \mathbb{R}^{d^2}$, we use I to indicate its representation as a matrix, i.e., $\mathrm{I} \in \mathbb{R}^{d \times d}$. Any image matrix can be written as $\mathrm{I} = [\mathrm{A}, \mathrm{B}]$ where the submatrices A and B are the

left and right sides of the image respectively.

Let $R : \mathbb{R}^{d^2} \rightarrow \mathbb{R}^{d^2}$ denote the reflection operator. It permutes the elements of an image such that if $\mathrm{I} = [\mathrm{A}, \mathrm{B}]$, then the matrix representation of RI is $[R\mathrm{B}, R\mathrm{A}]$. R is an involution, i.e., $R^2 I = I \quad \forall I$.

For simplicity, we consider a cell that has been exposed to just one template τ and its reflection $R\tau$. Consider the matrix representation of the (two-frame) video $V = (\tau, R\tau)^\intercal$. We are interested in the eigenvectors of the covariance matrix $V^\intercal V$. We show that $V^\intercal V$ commutes with the reflection R. Then, applying the fact that whenever two matrices commute they must have the same eigenvectors, we show that $V^\intercal V$'s eigenvectors must be symmetric or antisymmetric

First, we note that the eigenvectors of R, regarded as image matrices, are symmetric or antisymmetric about the vertical midline. That is, all the solutions of $Rv_\pm = \pm v_\pm$ are either of the form $v_+ = [\mathrm{A}, R\mathrm{A}]$ or $v_- = [\mathrm{A}, -R\mathrm{A}]$.

Next, we want to show that $RV^\intercal V = V^\intercal V R$.

Notice that we can write the covariance matrix as the sum of the outer products of V's columns:
$V^\intercal V = \tau\tau^\intercal + (R\tau)(R\tau)^\intercal$. Thus:

$$\begin{aligned} RV^\intercal V &= R\tau\tau^\intercal + RR\tau(R\tau)^\intercal \\ &= R\tau\tau^\intercal + \tau\tau^\intercal R \end{aligned}$$

and

$$\begin{aligned} V^\intercal V R &= \tau\tau^\intercal R + R\tau(R\tau)^\intercal R \\ &= \tau\tau^\intercal R + R\tau\tau^\intercal \end{aligned}$$

Therefore the covariance matrix commutes with the reflection; $RV^\intercal V = V^\intercal V R$. Thus they must have the same eigenvectors. Since the eigenvectors of R are symmetric and antisymmetric, the same must be true of the eigenvectors of $V^\intercal V$.

5.3.7 Properties of learned S-units as a function of their eigenvalue

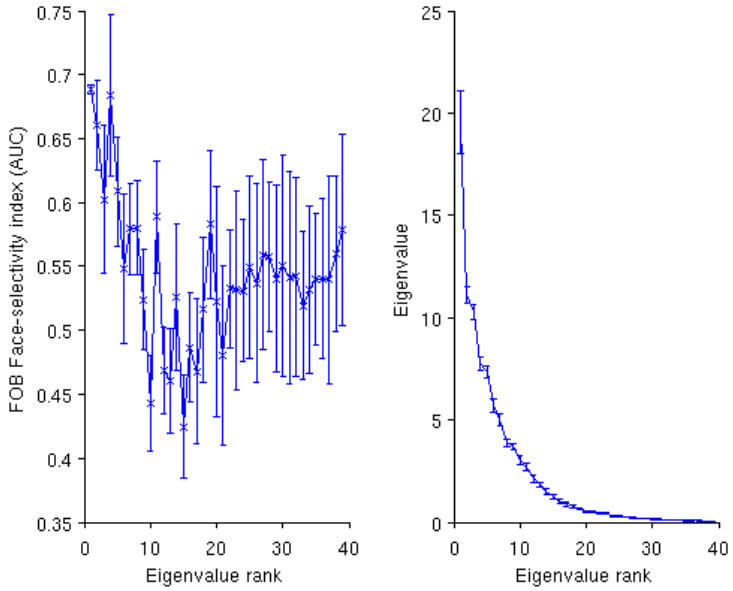

Figure 5-5: ***:****Left: Face-selectivity index as a function of eigenvalue rank.* ***Right:****Eigenvalue distribution.*

Only the first few eigenvectors give model units that are face-selective.

5.3.8 Yaw, pitch and roll rotations

A critical prediction of the theory is that AL cells should have equal responses to faces and their reflection over the vertical midline. The available data from [15] (figure 3, reprinted here) is in accord with this. However, a quirk of the way the data is presented in that paper obscures this fact somewhat. Example cell 4 appears to exhibit bimodal tuning curves for other rotations in addition to left-right angle rotation (yaw). Notice however, that the images for which this cell responds maximally are still reflections over the vertical midline of one another. Since a cell tuned to a symmetric template will always have a symmetric tuning curve [114], the response of this cell is in accord with the theory. Cell 1 appears to have a bimodal tuning curve over in-plane rotation, a result

which, if true, would go against the theory. However, a closer inspection of the data shows that this cell isn't really bimodal over in-plane rotation. Even in [15]-figure 3, you can see from the marginal tuning curve for "picture-plane angle" that this cell is barely modulated by in-plane rotation. To further confirm that this type of bimodal tuning curve does not appear in the available data, we manually inspected the analogous figure for all 57 cells recorded in the original study and did not find any with responses that were not in accord with the theory.

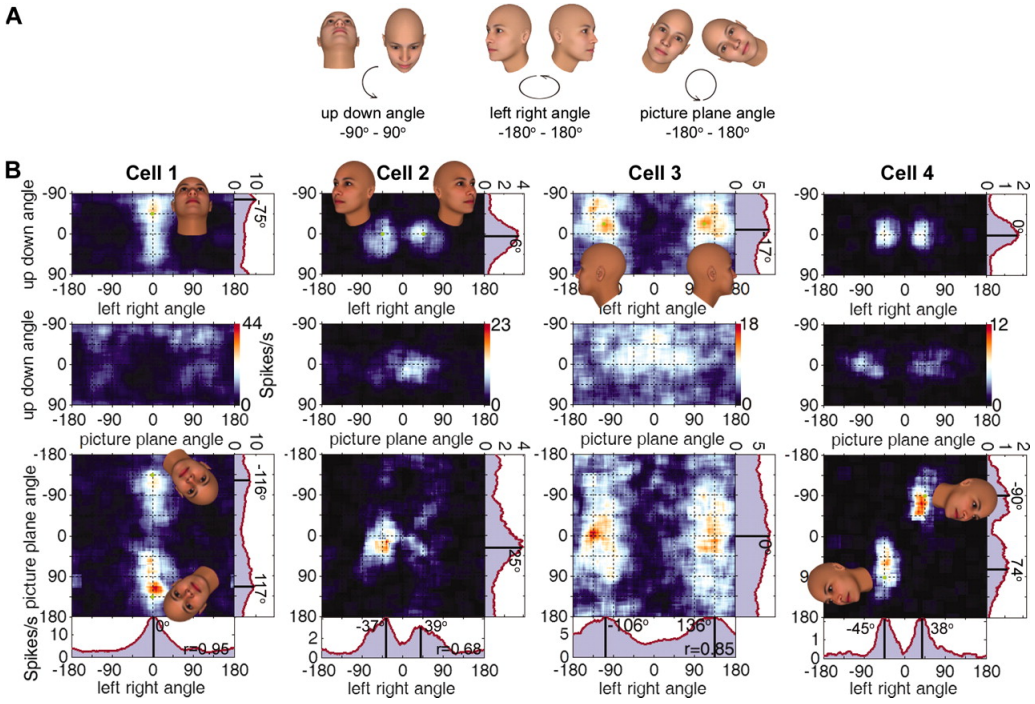

Figure 5-6: *From [15].*

Chapter 6

Discussion

6.1 Concerning modularity

6.1.1 Why should HW-modules for objects that transform similarly by arranged near one another on cortex?

The bulk of our argument has been concerned with why a recognition system ought to have dedicated circuitry to implement subordinate-level recognition for each object class. On its own, this does not say anything about the arrangement on cortex of the domain-specific circuitry. In principle, any HW-module could be anywhere, as long as the wiring all went to the right place. From this "only the synapses matter" perspective, there could be many, completely equivalent ventral streams with radically different spatial layouts. 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.

1. The wiring economy principle [115–118]

   (a) Intra-layer wiring economy — We argue by counterexample. Suppose the cells comprising HW-modules for two objects A and B are intermixed in one

cortical region. The only intra-area connections are from A-cells to A-cells or from B-cells to B-cells. In particular, they are from A's S-units to A's C-units, or from B's S-units to B's C-units. We assume—only in this section—that there is more than one C-unit for a given HW-module. Saying that A and B cells are intermixed implies that in between any two to-be-connected A-cells there might be a B-cell getting in the way. Inserting B-cells in between to-be-connected A-cells increases the length of wiring needed for the system to function. Thus, the wiring economy principle implies the regional segregation of different HW-modules. This is essentially the same argument used by Mitchison to explain the existence of multiple visual areas [117].

(b) Inter-layer wiring economy — If many cells all need to project to the same far-away place, then the length of wire required is smallest if both the presynaptic and postsynaptic cells are clustered.

2. Modulation / attention — There is substantial evidence that some forms of neuromodulation operate on local neighborhoods of cortex [119]. For example, neuromodulatory substances could diffuse beyond specific synapses and affect entire neighborhoods. Attention may operate through mechanisms of this sort—possibly involving acetylcholine or norepinephrine [120]. Given that the brain employs neuromodulatory mechanisms operating on local cortical neighborhoods, it makes sense that cells with similar functions would be grouped nearby one another. If it were not so, then such mechanisms could not be used effectively.

6.1.2 Other domain-specific regions besides faces and bodies

There are a few 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 is the macaque fruit area, recently discovered with high-field fMRI [80].

Lateral Occipital Complex (LOC) [101]

The present work implies that LOC is not really a dedicated region for general object processing, rather, it is a heterogeneous area of cortex containing many domain-specific 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 sometimes observed appearing in simulations.

Parahippocampal Place Area (PPA) [81]

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 nice, but would not also apply to other objects. 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 [121]. 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). Afterall, 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 [122, 123].

The Visual Word Form Area (VWFA) [82]

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 changes in the images received by our retina when a page is viewed with even a slightly different orientation. Even more severe, we are not confused by changing typefaces (font, weight, italics, etc). 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. Every writer draws their letters slightly differently, yet typical readers are invariant to all of this. Thus, VWFA is well-accounted for by the invariance hypothesis. Words are frequently-viewed stimuli which undergo class-specific transformations.

We note, however, that VWFA has some interesting other properties with respect to

our hypothesis. In particular, VWFA could not have evolved in the short (evolutionary) time since humans have been reading. Since we have not discussed how HW-modules could be learned here, this discussion is beyond the scope of the present paper (but see this technical report [36] for additional discussion).

6.1.3 Psychophysics of viewpoint-tolerance

There is a large literature emphasizing the view-dependence of human vision when tested on subordinate level tasks with unfamiliar examples—e.g. [41, 66, 89]. Our proposal 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 [93, 124–126], and before differences between the view-tolerance achieved on basic-level and subordinate-level tasks were fully appreciated [42, 127, 128].

The view-dependence observed in experiments with novel faces [66, 129] is fully consistent with the predictions of our theory. The 3D structure of faces does not vary wildly within the class, but there is certainly still some significant variation. It is this variability in 3D structure within the class that is the source of the imperfect performance. Many psychophysical experiments on viewpoint invariance were performed with synthetic "wire" objects defined entirely by their 3D structure. We found that they were by far, the least nice (lowest $\overline{\psi}$) objects we tested (see supplementary table 1 above). Thus our theory predicts particularly weak performance on viewpoint-tolerance tasks with (novel examples of) these stimuli and that is precisely what is observed [88].

An interesting experiment by Tarr and Gauthier is relevant here [42]. They found that learned viewpoint-dependent mechanisms could generalize across members of a homogenous object class. They tested both homogenous block-like objects, and another class 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 theory. Interestingly, in light of our results here, it would be natural to predict better

within-class generalization of learned "view-dependent mechanisms" (to use Tarr and Gauthier's vocabulary) for object classes with higher niceness. That is, transformation compatibility, not visual similarity per se, may be the factor influencing the extent of within-class generalization of learned view-tolerance.

6.2 Concerning learning, plasticity, and mirror-symmetric tuning

We showed in chapter 5 that a biologically-plausible learning rule—related to Oja's and Foldiak's rules—can be used to learn HW-modules from video. The HW-modules so-learned have a few different properties than the standard view-based/memory-based modules. In particular, they can be seen as *compressed* memory-based models; their templates are given by solutions to an eigenvector equation we called the cortical equation. We studied the results of applying this rule in simplified settings with videos that depict only a single object undergoing a single transformation (chapters 3 and 5). In those cases, we showed that the compressed modules have properties resembling ventral stream neurons. When we solved the cortical equation with parameters modeling early visual areas (small/short spatiotemporal timescale) we obtained Gabor-like tuning. Solving the cortical equation with parameters modeling a high-level face/depth-rotation specific area (large/long spatiotemporal timescale) produces model units with mirror-symmetric face orientation tuning curves similar to those observed in the macaque face patch AL.

In my opinion, the most interesting aspect of the work on learning HW-modules is the future research directions it opens up. Despite the hype about "big data", the next fundamental challenge for machine learning and artificial intelligence is understanding how humans learn from very small numbers of examples, and replicating that behavior in a machine. The learning mechanism proposed here is a form of representation learning. It is a way of learning to compute representations for new data that are invariant to the transformations undergone by the old data (in the case of group transformations, these are exactly the same). In object recognition the natural use for HW-modules is invariance to identity-preserving image transformations. However, that is not the only setting where HW-modules, and the proposed learning rule, can be applied. In the standard machine learning supervised categorization setting, HW-modules with S-units tuned to examples from one class can be used to classify new inputs. In that case, building an HW-module can be seen as a particular method of training a classifier—a highly efficient method at that. Learning a compressed HW-module would be training another classifier, possibly related to spectral regularization methods [130]. The latter method has a convenient interpretation as an incremental (online as opposed to batch) algorithm via Oja's rule, or one of its variants. We have already applied some of these "supervised" HW-modules and obtained good results on unconstrained face recognition benchmarks [56]. Notice that I'm putting the scare-quotes around the word "supervised" for a specific reason. I don't actually think that the brain needs supervision to learn how to recognize faces. Rather, exposure to faces, and the operation of the right unsupervised rule leads to the emergence of representations that are invariant to face-specific transformations. One possible way to study this further could be to move away from the simplified settings we have so far considered and apply these learning rules to natural (or semi-natural computer-generated) video. Many new questions arise in that setting: e.g., how do you prevent all the HW-modules from learning the same thing as one another? How should the spatiotemporal apertures and layers be arranged? What about gating—and splitting off during learning—of class-specific hierarchies? This research direction is really wide open at present—the limits of what hierarchies of HW-modules can learn to compute is still far from clear. There is much

more work to be done!

I conclude with a conjecture. Extensions of the general theory of invariance beyond vision, and beyond the ventral stream, are possible. Since domain-specific regions for abstract concepts like “number” and various aspects of language apparently exist in other brain areas [131], it is interesting to consider the hypothesis that these too may be explainable in this framework. One possible connection could be through Tarski’s characterization of logical objects by their invariance under *arbitrary* transformations [132].

Appendix A

Subtasks of Unconstrained Face Recognition

Unconstrained face recognition remains a challenging computer vision problem despite recent exceptionally high results ($\sim 95\%$ accuracy) on the current gold standard evaluation dataset: Labeled Faces in the Wild (LFW) [64, 133]. We offer a decomposition of the unconstrained problem into subtasks based on the idea that invariance to identity-preserving transformations is the crux of recognition. Each of the subtasks in the *Subtasks of Unconstrained Face Recognition* (SUFR) challenge consists of a same-different face-matching problem on a set of 400 individual synthetic faces rendered so as to isolate a specific transformation or set of transformations. We characterized the performance of 9 different models (8 previously published) on each of the subtasks. One notable finding was that the HMAX-C2 feature was not nearly as clutter-resistant as had been suggested by previous publications [7, 134]. Next we considered LFW and argued that it is too easy of a task to continue to be regarded as a measure of progress on unconstrained face recognition. In particular, strong performance on LFW requires almost no invariance, yet it cannot be considered a fair approximation of the outcome of a detection$\rightarrow$alignment pipeline since it does not contain the kinds of variability that realistic alignment systems produce when working on non-frontal faces. We offer a new, more difficult, natural image dataset: SUFR-in-the-Wild

(SUFR-W), which we created using a protocol that was similar to LFW, but with a few differences designed to produce more need for transformation invariance. We present baseline results for eight different face recognition systems on the new dataset and argue that it is time to retire LFW and move on to more difficult evaluations for unconstrained face recognition.

A.1 Introduction

Current approaches to face recognition perform best on well-posed photographs taken for identification purposes, e.g., passport photos. However, in the real world, images of faces undergo many transformations—including aging, pose, illumination, expression, and many more. Not only do transformations degrade the performance of current algorithms, but in many cases they are known to lead to their catastrophic failure [135, 136].

The computer vision and biometrics communities have responded to this challenge by shifting their focus to unconstrained benchmark datasets, of which Labeled Faces in the Wild (LFW) is generally considered to be the gold standard [64]. LFW and similar datasets (e.g., PubFig83) consist of publicly available images of celebrities gathered from the internet and thus contain considerable variability.

The state-of-the-art on LFW has steadily improved in recent years to the point that it now arguably rivals human performance (on same-different matching of unfamiliar faces). At the time of writing, the best LFW performance is above 95% [133]. However, we argue in this paper, there are several reasons that a declaration of victory over unconstrained face recognition remains premature.

1. The strong performance achieved on Labeled Faces in the Wild does not transfer to another, ostensibly quite similar, dataset we gathered.

2. The failure modes of state-of-the-art algorithms remain unclear. Moreover, when an algorithm does not perform well on an unconstrained test like LFW, it is not clear what aspect of the task is responsible.

3. Another goal is to understand the brain's solution to the unconstrained face recognition problem. In the Visual Psychophysics and Cognitive Neuroscience litera-

ture there is a wealth of available information concerning the robustness of human vision with respect to specific transformations, e.g, [66, 137]. This data is typically gathered in highly controlled laboratory settings with one transformation varied at a time. Unless artificial systems are tested in comparable settings then there is no way to connect to this large body of previous work.

In this paper, we argue that in order to make further progress, it is necessary to simultaneously consider unconstrained face recognition along with its component subtasks. To that end, we contribute a collection of synthetic datasets (produced using 3D graphics) which, taken together, constitute a (partial) decomposition of unconstrained face recognition into its component subtasks. Our parsing of the full problem into subtasks is based on the premise that transformation invariance is the crux of recognition [36, 87]. We also gathered a new unconstrained dataset, similar to LFW (publicly available images on the Internet), but apparently more difficult. The entire collection of new datasets is available to researchers[1].

A.2 Subtasks

Our decomposition of unconstrained face recognition into subtasks is based on the idea that invariance to transformations is the main computational problem of recognition. The subtasks can be used to test face recognition systems. Unlike LFW, and similar datasets for which only a single accuracy score is measured, testing on all the subtasks gives a detailed analysis in terms of which transformations a system handles well and which cause it to fail.

The Subtasks of Unconstrained Face Recognition (SUFR) challenge is a collection of datasets which we call subtasks. Each subtask was designed to test specific aspects of the unconstrained face pair-matching (same-different) task. There are 400 individuals in each subtask. The total numbers of images range from $2,000$ for some of the smaller subtasks, to $10,000$ for some of the larger interaction tasks (tests with two transformations applied simultaneously). Each image is 512×512 pixels and in color.

[1] It can be downloaded from `cbcl.mit.edu`.

Since our goal in creating these datasets was precise control of transformation parameters, we employed 3D graphics software to synthesize the images. In section A.3.1 we also describe a separate component of the challenge which uses natural images: SUFR-W.

The 400 textured head models were randomly generated using FaceGen [103] and rendered onto a transparent background with Blender [102] using the CYCLES ray tracing engine. Most of the transformations required 3D information, e.g., rotation in depth and simulated movement of the illumination source. These transformations were applied with Blender. In other cases, images were transformed by explicitly specifying an affine matrix and using Matlab's image processing toolbox.

The SUFR challenge can be divided up in different ways. The "core" of the challenge is a set of six datasets which test transformation invariance directly. They consist of images of faces on a uniform black background. Another set of subtasks are concerned with transformation invariance in the presence of background clutter. Each image has a different randomly chosen natural scene or semi-structured random noise image in the background. Several subtasks are suitable for studying robustness to occlusion. Strong performance on these tasks requires invariance to whether or not a face is wearing sunglasses. Finally, there are also interaction tests. It is possible for a face recognition system to employ methods that successfully ensure invariance to any single transformation, but fail in combination. The interaction tests could quickly diagnose such issues. The full list of subtask datasets and benchmark results (without the random noise background sets for space reasons) is in table A.1.

Testing face recognition algorithms on all the SUFR datasets yields a lot of information. However, it should be noted that SUFR is still only a partial decomposition of the unconstrained face recognition problem. In general, it would have to include transformations that are quite difficult to parametrize, e.g., facial expressions and aging. Thus our parsing of the full task remains somewhat incomplete since it only contains the transformations which we were able to simulate using 3D graphics software. Nevertheless, the SUFR benchmark contains many tests which are quite difficult for recent face recognition systems.

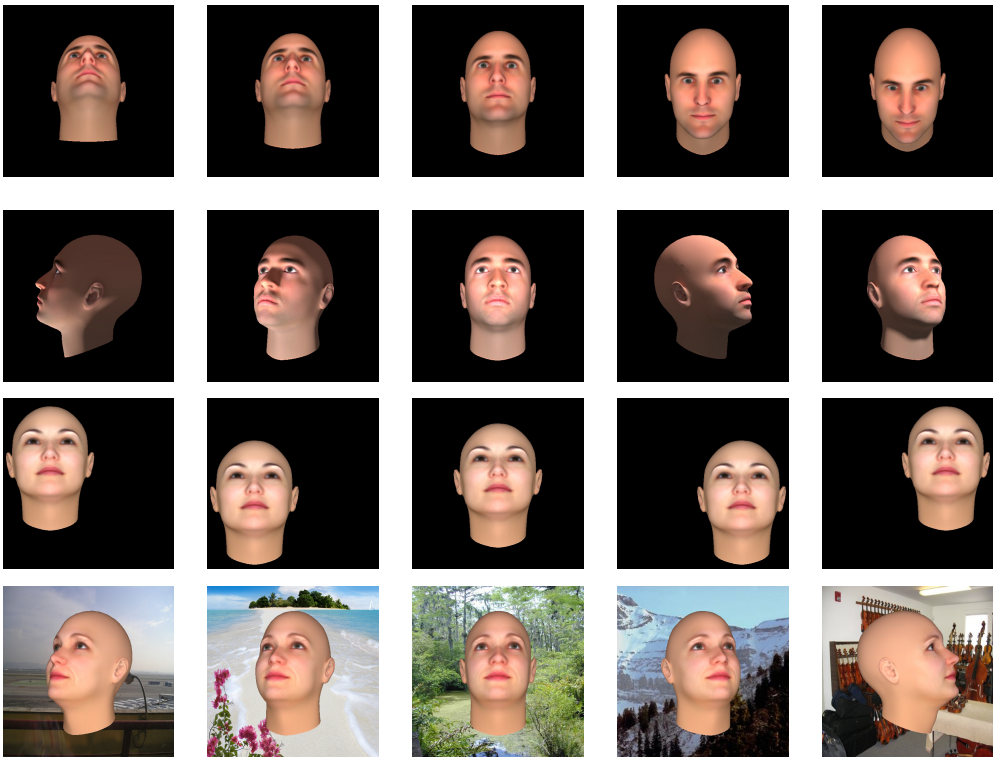

Figure A-1: *Example images.*

A.2.1 Performance of benchmark face recognition models

The intended use of the SUFR datasets is same-different matching of unfamiliar individuals (never seen during the training phase). This problem is sometimes called face-verification. It is identical to the standard procedure used with LFW. Unless mentioned otherwise, each test was performed by training a Support Vector Machine (SVM) using the difference between the feature representations of the two images to be compared. 4000 image pairs were used for training and 4000 independent pairs for testing.

The SUFR benchmark results in table A.1 include all nine models we tested. However, some care in interpretation is needed since they are not all directly comparable with one another. For example, some entries in table A.1 correspond to testing concatenated vectors of local descriptors for their translation invariance. Obviously, they are not translation invariant—they were never intended to be.

Local descriptors

Many computer vision features (e.g. Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP)) are extracted independently on a block-by-block basis. That is, first each image is subdivided into many relatively small blocks, then each is fed into a "feature extraction blackbox" which returns a vector of feature values. Vectors from all the blocks are concatenated to represent the entire image. Within this category, we tested Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Local Phase Quantization (LPQ). *Note:* Many of these features could be used as components of Global methods using bag of words or spatial pyramid approaches. We list them as "local" since their particular variant tested here was local.

Histograms of Oriented Gradients (HOG)

Originally proposed by **(author?)** [138], our experiments are based on the variant proposed by **(author?)** [139]. The image was divided into blocks. For each one, a histogram of gradient orientations for each pixel is accumulated. The histogram of each block is then normalized with respect to neighboring blocks. We used an open source implementation from the VLFeat library [140].

Local Binary Patterns (LBP)

LBP [141] and its generalizations to three-patch-LBP, four-patch-LBP and Local Quantized Patterns have been shown to be powerful representations for face recognition with LFW [142–144]. These methods work by thresholding the pixel intensities in a small region surrounding a central pixel and treating the resulting pattern as a binary number. As in HOG, histograms of local descriptors are accumulated in non-overlapping blocks. We used the implementation from VLFeat [140].

Local Phase Quantization (LPQ)

LPQ [145] is a blur-insensitive feature computed by quantizing the Fourier transform phase in local neighborhoods. Variants of LPQ were previously shown to outperform LBP on several datasets including LFW [146]. We used an implementation provided by

the author.

Features inspired by primary visual cortex

Hierarchical Model and X — C1 (HMAX-C1)

HMAX is a (partial) model of the primate ventral stream [3], the part of cortex that is believed to be involved in object recognition. The elements of its C1 layer model complex cells in primary visual cortex (V1). We used the open source "CVPR06" implementation of HMAX which is distributed with the CNS simulation system [147].

V1-like model (V1-like)

V1-like features are another family of low-level features intended to model the output of primary visual cortex [135]. Variants of V1-like features were shown to be effective in various object and face recognition tasks [96, 135]. In all of our experiments, we used V1-like(A)—the best performing variant according to [96]. We used an implementation provided by the author. Following their testing procedure, we reduced the dimensionality of the V1-like features by PCA[2] [148].

Global features

Hierarchical Model and X — C2 (HMAX-C2)

Another layer of HMAX. It was developed as a model for regions involved in later stages of ventral stream visual processing beyond primary visual cortex. We used the open source "PNAS" implementation of HMAX from CNS [147]. This version corresponds to the "C2b" layer of [52].

Scale-Invariant Feature Transform + Bag of Words or Spatial Pyramid (SIFT-BoW and SIFT-Pyr)

The Scale-invariant feature transform (or SIFT) [149] is performed on a point-by-point basis. Canonically, 128 dimensional features can be extracted from a keypoint, but one cannot directly use it for classification. A common practice is to use a Bag-

[2]Due to the large size of the features (86,400 per image) we only used 1,000 random training samples (out of 4,000) to compute principal components.

of-words (BoW) or spatial pyramid representation (Pyr), which treats each keypoint as a visual word and ignore its spacial location in the whole image (BoW) or each block (Pyr). A histogram of all visual words is computed as the final features. We used k-means clustering to quantize these visual words into 1024 clusters producing a final feature size of 1024 (BoW) or N*1024 (Pyr), where N is the number of blocks in the spatial pyramid. The open source implementation is provided by [150].

An alignment-based system
SIFT-RANSAC→Warping→HOG features

We developed and tested the following pipeline—SIFT-RANSAC→Warping→HOG features. The idea is: given a pair of test images, warp the first image, A, to match the other image, B. If the warping is successful then A could be aligned with B and substantial affine transformations discounted. Since many other transformations are approximately affine (e.g. small yaw rotations) it is possible that this approach may also be successful in those cases. We implemented the common SIFT-RANSAC algorithm that is usually used for panoramic photo stitching. Then we extracted HOG features from image B and the warped image A. After that, we followed the same testing process as with the HOG features.

The SUFR benchmark clusters models by type
We used multidimensional scaling (MDS) to visualize the similarities between the pattern of results obtained with each feature set (fig. A-2). Distance between points in the scatter plot corresponds to the Euclidean distance between each model's vector of accuracy values on the "core SUFR" subset: all single transformation subtasks with a uniform background. It shows that the feature types can be distinguished from one another by their pattern of SUFR results. Unsurprisingly, one MDS dimension appears to represent "globalness", HMAX-C2, the two SIFT-based models, and the RANSAC-HOG system are located at its extremes. The more local models inspired by primary visual cortex: HMAX-C1 and Pinto's V1-like model also cluster closer to one another than to other models, though interestingly, they are farther apart than we expected. A more surprising finding was that HOG, LPQ, and LBP all had quite similar patterns of results on

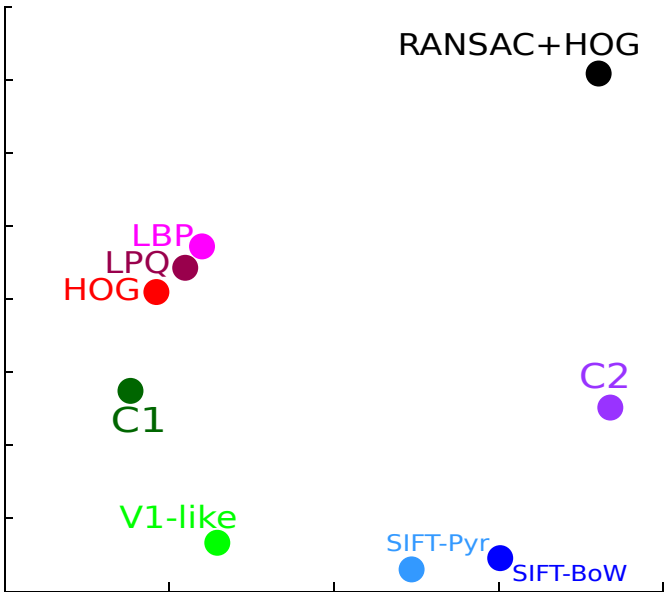

Figure A-2: Multidimensional scaling based on the vector of performances on cure SUFR. Distance in the scatter point corresponds to the Euclidean distance between each model's vector of accuracies on the core SUFR tests.

the SUFR benchmark and all were relatively different from the local biologically-inspired features. As expected, the RANSAC-HOG system is isolated and far from other features. It works very well for all the affine transformations (even with background). But for non-affine transformations, it is fairly unstable and largely compromised, the same reason it is not applicable to real-world data.

Disrupting translation invariance with cluttered backgrounds

HMAX-C2 and SIFT-Bag-of-Words performed nearly perfectly on the tests of translation invariance without background clutter. However, both failed the same test in the presence of natural image clutter. This result was surprising since there are at least two previous reports in the literature that HMAX-C2 was translation-invariant on tasks with cluttered backgrounds [7, 134].

(author?) [7] tested translation-invariant face pair-matching with and without background clutter. They reported that there was very little loss of accuracy due to clutter. However, it is likely that the clutter they used was too noise-like and not similar enough to the target class (natural faces). We observed that random semi-structured noise

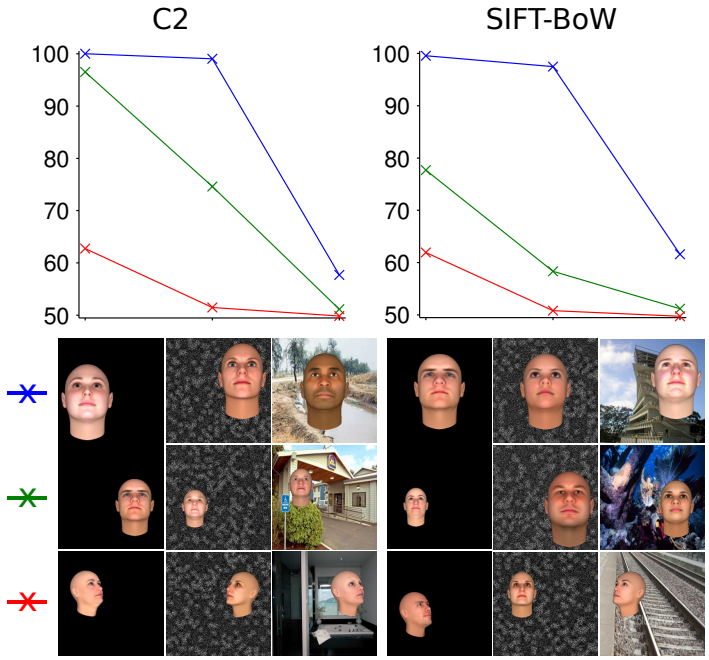

Figure A-3: Top diagrams: Accuracy curves of C2 and SIFT-BoW over different transformations and background types (Blue: translation, Green: translation + scaling, Red:translation + yaw rotation). Y axis is verification accuracy in percentage. X axis is background type. 1 = no background. 2 = noise. 3 = natural images. Bottom row shows the example images used for the three curves, respectively.

backgrounds do not have much effect on translation invariance for either HMAX-C2 or SIFT-BoW (fig. A-3).

(author?) [134] followed a similar approach to ours. They also generated datasets of transforming objects using 3D graphics. However, they studied a basic level categorization task: cars vs. airplanes. They found that HMAX C2's performance was unaffected by translation over natural clutter. It is possible that this result was due to a difference between subordinate level face matching and their basic level task. But there were many other differences between the two studies that may also have been responsible. We also investigated a pure background-invariance task which was trivially easy for the local features and found that C2 and the SIFT-BoW method were quite disrupted by very small amounts of clutter—even when no translation invariance is necessary (fig. A-4).

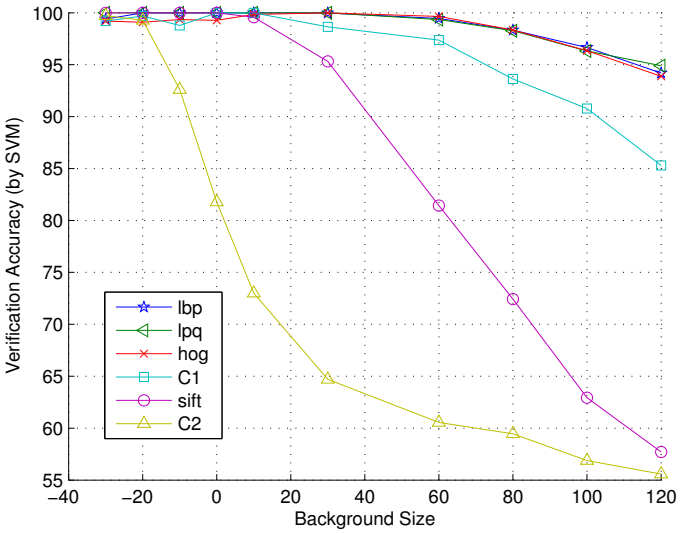

Figure A-4: Performances of different models over different background sizes. It shows that global features (e.g., C2, SIFT) are much less tolerant of clutters, adding even a small amount of background lower their performances significantly.

A.3 Face recognition in the wild

If you accept the premise that transformation invariance is the crux of visual recognition then performance on the subtasks ought to be a good predictor of performance on the unconstrained task. However, if the meaning of "the unconstrained task" is "Labeled Faces in the Wild", this turns out not to be true. Figure A-6 shows that many of the models we tested actually perform better on LFW than they do on most of the subtasks. How can this be?

It turns out that LFW doesn't really require substantial invariance to many of the transformations that the SUFR datasets were designed to test. The creators of LFW filtered its set of candidate images by the Viola-Jones face detector [151] which, for the most part, only detects nearly frontal faces. Thus LFW contains hardly any rotation in depth. Also, the faces are all centered and roughly the same size so translation and scale invariance are also unnecessary.

Table A.1: Subtasks of Unconstrained Face Recognition benchmark results (% correct).

Core	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
translation	52.8	**99.6**	53.0	55.0	55.9	**98.0**	**89.6**	69.6	**93.7**
scaling	61.7	**87.5**	61.7	61.0	62.7	64.7	63.7	55.3	**80.5**
in-plane rotation	61.4	**85.9**	71.3	79.3	71.2	77.9	71.5	63.1	**99.4**
pitch rotation	79.5	**90.0**	79.8	**84.1**	76.5	79.7	75.9	70.5	76.2
yaw rotation	57.1	**70.8**	58.6	64.8	60.3	67.1	63.1	59.8	55.1
illumination	**96.0**	**94.6**	**93.2**	**92.5**	**87.2**	**93.1**	**95.5**	**96.3**	71.7
Core + clutter	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
translation	55.5	57.7	57.1	57.6	57.3	61.6	55.5	49.6	**97.1**
scaling	49.6	48.4	53.3	53.5	52.6	51.0	52.2	49.4	**89.5**
in-plane rotation	54.6	50.7	54.5	60.2	55.7	51.3	51.0	53.2	**96.6**
pitch rotation	54.1	52.5	54.5	60.1	55.9	51.0	52.7	55.4	**68.2**
yaw rotation	49.6	48.5	50.7	52.2	51.4	49.7	49.8	50.5	52.7
illumination	56.0	49.6	**67.0**	62.9	60.6	50.1	50.6	58.2	54.7
Interactions	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
translation + scaling	53.5	**96.5**	53.0	53.2	53.3	77.7	67.6	51.5	**84.5**
translation + in-plane rotation	53.4	**87.1**	53.3	53.3	52.5	79.2	57.6	51.5	**91.8**
translation + yaw rotation	50.5	**62.7**	51.3	51.2	51.3	**62.0**	52.1	51.3	51.7
yaw rotation + illumination	56.5	58.5	52.6	54.2	54.9	59.3	57.1	57.4	52.7
Occlusion	C1	C2	HOG	LBP	LPQ	SIFT-BoW	SIFT-Pyr	V1-like	RANSAC+HOG
sunglasses + pitch rotation	76.6	69.5	79.7	**84.5**	77.6	75.8	73.5	64.2	63.6
sunglasses + yaw rotation	57.0	50.0	59.8	**69.3**	61.3	**67.9**	63.6	59.5	54.8

A.3.1 SUFR in the Wild (SUFR-W)

In order to address these shortcomings of LFW, we created a new "unconstrained" natural image dataset using a very similar protocol to the one used by the creators of LFW. The new dataset, which we call SUFR-in-the-Wild (SUFR-W), is similar in size to LFW. It contains 13,661 images, slightly more than LFW's 13,233. While LFW contains a small number of images per person and a large number of people (5749 individuals), SUFR-W contains a much larger number of images of exactly 400 people (picked for uniformity with the synthetic SUFR datasets). See figure A-5 for example SUFR-W images.

We gathered the images for SUFR-W using Google images. In order to avoid the

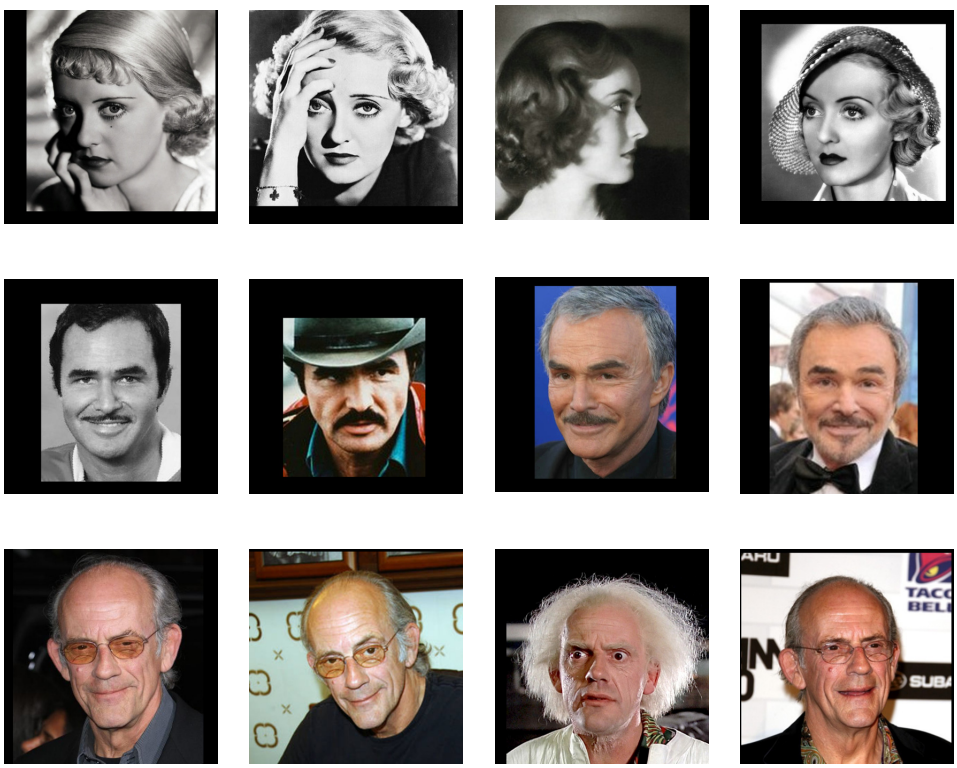

Figure A-5: Example images in the SUFR-in-the-Wild dataset (SUFR-W). Top row: Bette Davis. Middle row: Burt Reynolds. Bottom row: Christopher Lloyd. The degree of alignment shown here is typical for the dataset. Profile faces as in the top row are rare.

same Viola-Jones filtering issue that prevented LFW from containing non-frontal faces, we did the following: First we manually eliminated all the images for each name that did not have a single isolated face, were not the correct person, or were too low resolution. Next, to prevent the dataset from being too difficult, we ran the **(author?)** [152] face detection and landmark localization method. This method works particularly well with rotations in depth. It managed to detect all but ~ 30 of the candidate faces (which we then removed). To introduce some additional difficulty, but not too much, we allowed the **(author?)** [152] system to attempt to align the images based on the landmarks it localized. However, it frequently failed to achieve a good alignment. Many of the faces (but not too many) remain clearly misaligned. Since we performed no further alignment, all these images are still misaligned in the final dataset.

SUFR-W contains none of the same individuals as LFW so it is straightforward to

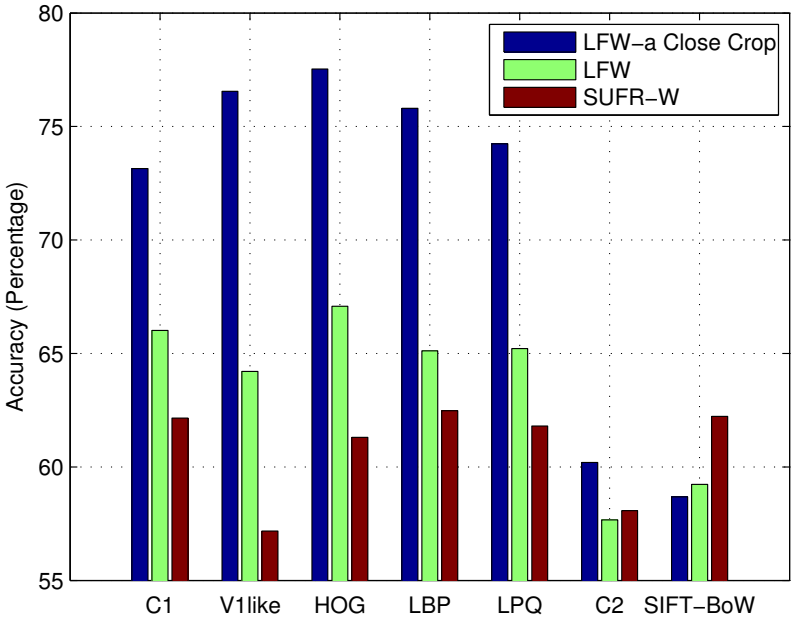

Figure A-6: Results on natural image tasks (LFW-a closely cropped, LFW original and SUFR-W). The x axis is type of features. All the results are from our experiments, except that the LFW V1like is from [148] and LFW-a close crop V1like is reported in [96]. Our attempts to replicate these were stymied by a lack of computational resources

conduct experiments that train on one dataset and test on the other. As an unintended consequence of this, since so many celebrities are already in LFW, we had to look farther afield to find the individuals for SUFR-W. Many of them are actors and politicians who were active in the first half of the 20th century. Since these individuals are older today, we found that SUFR-W has considerably more variation in age than LFW. Of course, one unfortunate bias is that age is very clearly correlated with photography style (e.g. 'younger' implies 'probably black and white'). This is not a problem for the same-different matching task; though it does mean that successful algorithms will need to be reasonably tolerant of "the aging transformation".

While the systems we tested are not quite at the state-of-the-art, it is clear from the difference in performance between LFW and SUFR-W that the latter is a considerably more difficult dataset (fig. A-6). At the same time, it is also clear that it is not so difficult that it cannot be used productively to guide future research.

Figure A-7: Top: Three typical images from LFW. Bottom: The same three images in LFW-a.

Upon seeing figure A-7, a colleague of ours remarked that the images in the *bottom* row are the ones for the task of face *recognition*. Depending on what part of the community you come from, that statement will either be obviously true or completely absurd.

Of the 123 papers indexed by Google Scholar that report results on LFW, at least 95 of them actually used a different, even more tightly aligned version[3]. Most of these paper (at least 58 of them) used LFW-a, a version of LFW which was *very* finely aligned with a commercial software package [143]. The vast majority of papers using LFW-a crop all the images to an extremely tight, fixed, bounding box like the one shown in A-7.

Even the relatively simple features we tested here are improved by up to 10% by using (cropped) LFW-a (fig. A-6). Similar results have been reported before (e.g. **(author?)** [143]).

The argument in favor of taking results on the tightly cropped LFW-a test as a proxy for performance on unconstrained face recognition appeals to the detection $\rightarrow$ alignment $\rightarrow$ recognition (DAR) pipeline. In that framework, recognition is only the last step in a process through which transformations have already been discounted. It is ac-

[3]There were 9 papers that reported results on both and 23 papers for which we were unable to determine which dataset was used.

ceptable to focus on a dataset containing hardly any transformations since normalizing those was already supposed to have been accomplished at earlier stages. However, there are several reasons not to take this argument at face value.

1. At best, the DAR framework guarantees that recognition systems will receive data that is as well-normalized as detection and alignment systems can deliver within application demands (e.g. processing time or cost). The creators of LFW referred to this issue when they wrote

"every face image in our database is the output of the Viola-Jones face detection algorithm. The motivation for this is as follows. If one can develop a face alignment algorithm (and subsequent recognition algorithm) that works directly on LFW, then it is likely to also work well in an end-to-end system that uses the Viola-Jones detector as a first step." [64].

This view of LFW is very conservative with respect to its implications for the full unconstrained face recognition problem. In this vein, the honest interpretation of the fact that the state-of-the-art on LFW-a is now 95% is: Consider the algorithm that first runs Viola-Jones (missing all the non-frontal faces), then has humans manually remove false positives, then passes the remaining images to the commercial system used to create LFW-a, and finally, then runs the best performing system on LFW-a. 5% of this algorithm's error rate would be attributed to the last step.

2. Within the DAR framework, a more fair natural image recognition test along the lines of LFW would, at least, have to include the kinds of images obtained by the errors of the previous stages. At least, these images should be included if the results are to be understood as measuring progress on unconstrained face recognition. Even if one expects to have relatively strong detection and alignment in the pipeline, it is still desirable for the last step to tolerate transformations. This allows the recognition system to "rescue" some alignment errors. It introduces redundancy into the system and prevents alignment from being a single point of failure.

3. It is interesting to consider to what extent, if any, the DAR framework is applicable to the brain's method of recognizing faces. Eye movements serve to approximately align images across time. However, numerous studies have shown that the brain's

visual system is surprisingly tolerant of transformations, even when the images are flashed more quickly than the eyes can move [18]. One interpretation is that the brain's visual system has two operating modes. One mode is faster and more automatic; it does not involve eye movements. The other mode operates more slowly, engages specific task-related information, and employs eye movements for alignment.

A.4 Conclusion

It has long been appreciated that the development of appropriate recognition tests to isolate subproblems is essential to advancing computer vision. Notable datasets in this tradition include the Face Recognition Grand Challenge (FRGC) [153] and Multi-PIE datasets [154]. Approaches based on synthetic data have fallen out of favor in recent years. While synthetic tests clearly have limitations: the variability within the class of synthetic faces does not approach that of natural faces. Tests with synthetic data also have numerous advantages. In particular, appearance transformations can be specified with a level of detail that could never be obtained in a dataset of natural photographs. Very large synthetic datasets can be created with no extra cost, in the case of the SUFR challenge, it was simple to include tests that address interaction effects between transformations. This could not have been done in a set of natural photographs without a costly investment.

We advocate an approach that combines tests on unconstrained natural image datasets like Labeled Faces in the Wild with detailed testing of particular subtasks. However, the results presented here, and (much more so) the work of **(author?)** [133]—the creators of the current (95%) state-of-the-art system for LFW—argue that LFW may simply be too easy of a dataset to guide future progress. We suggested that the next generation of datasets ought to focus more on the problem of transformations. To that end, we are making the new SUFR-W dataset, as well as the complete set of synthetic datasets, available to interested researchers.

Bibliography

[1] F. Anselmi, J. Z. Leibo, J. Mutch, L. Rosasco, A. Tacchetti, and T. Poggio, "Unsupervised Learning of Invariant Representations in Hierarchical Architectures (and in Visual Cortex)," *Under review*, 2013. 2, 12, 18, 20, 27, 74, 80, 81, 82, 83

[2] D. Marr and T. Poggio, "From understanding computation to understanding neural circuitry," *AIM-357*, 1976. 2, 22

[3] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, Nov. 1999. 10, 15, 21, 22, 26, 27, 47, 64, 81, 111

[4] 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*, vol. 29, no. 3, pp. 411–426, 2007. 10, 15, 21, 26, 27, 43, 47, 48, 51, 52, 56, 64, 81

[5] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," *IEEE International Conference on Computer Vision*, pp. 2146–2153, 2009. 10, 21, 44

[6] 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. 10, 21, 44

[7] J. Z. Leibo, J. Mutch, L. Rosasco, S. Ullman, and T. Poggio, "Learning Generic Invariances in Object Recognition: Translation and Scale," *MIT-CSAIL-TR-2010-061, CBCL-294*, 2010. 10, 21, 25, 48, 57, 65, 92, 105, 113

[8] J. Z. Leibo, J. Mutch, S. Ullman, and T. Poggio, "From primal templates to invariant recognition," *MIT-CSAIL-TR-2010-057, CBCL-293*, 2010. 10, 25

[9] D. Cox, P. Meier, N. Oertelt, and J. J. DiCarlo, "'Breaking' position-invariant object

recognition," *Nature Neuroscience*, vol. 8, no. 9, pp. 1145–1147, 2005. 10, 17, 24, 46, 54, 55, 56, 57, 64, 85

[10] N. Li and J. J. DiCarlo, "Unsupervised natural experience rapidly alters invariant object representation in visual cortex.," *Science*, vol. 321, pp. 1502–7, Sept. 2008. 10, 17, 24, 46, 52, 53, 54, 55, 56, 64, 85

[11] J. Z. Leibo, J. Mutch, and T. Poggio, "Why The Brain Separates Face Recognition From Object Recognition," in *Advances in Neural Information Processing Systems (NIPS)*, (Granada, Spain), 2011. 11, 57, 65, 81, 83, 84, 92

[12] N. Kanwisher, J. McDermott, and M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *The Journal of Neuroscience*, vol. 17, no. 11, p. 4302, 1997. 11, 60, 76

[13] D. Tsao, W. A. Freiwald, T. Knutsen, J. Mandeville, and R. Tootell, "Faces and objects in macaque cerebral cortex," *Nature Neuroscience*, vol. 6, no. 9, pp. 989–995, 2003. 11, 60, 61

[14] D. Tsao, W. A. Freiwald, R. Tootell, and M. Livingstone, "A cortical region consisting entirely of face-selective cells," *Science*, vol. 311, no. 5761, p. 670, 2006. 11, 60, 61, 92

[15] W. A. Freiwald and D. Tsao, "Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System," *Science*, vol. 330, no. 6005, p. 845, 2010. 11, 61, 62, 81, 83, 91, 92, 96, 97

[16] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001. 12

[17] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, 1996. 12, 62

[18] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo, "Fast Readout of Object Identity from Macaque Inferior Temporal Cortex," *Science*, vol. 310, pp. 863–866, Nov. 2005. 12, 26, 33, 121

[19] L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio, "The dynamics of invariant object recognition in the human visual system," *Submitted*, 2013. 12, 33

[20] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943. 12

[21] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, p. 106, 1962. 13, 14, 15, 47, 56, 64, 81, 82

[22] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 284–299, 1985. 14, 86

[23] C. G. Gross, D. Bender, and C. Rocha-Miranda, "Visual Receptive Fields of Neurons in Inferotemporal Cortex of Monkey," *Science*, vol. 166, pp. 1303–1306, 1969. 15

[24] C. G. Gross, C. Rocha-Miranda, and D. Bender, "Visual properties of neurons in inferotemporal cortex of the macaque," *Journal of Neurophysiology*, vol. 35, no. 2, pp. 96–111, 1972. 15

[25] N. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Current Biology*, vol. 5, no. 5, pp. 552–563, 1995. 15, 23, 25, 32, 66, 75, 76

[26] J. J. DiCarlo and J. Maunsell, "Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position," *Journal of Neurophysiology*, vol. 89, no. 6, p. 3264, 2003. 15, 25, 32

[27] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, Apr. 1980. 15, 21, 26, 47, 64, 81

[28] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, pp. 255–258, 1995. 15, 26, 64, 80, 81

[29] J. Mutch and D. Lowe, "Multiclass object recognition with sparse, localized features," *Computer Vision and Pattern Recognition 2006*, vol. 1, pp. 11–18, 2006. 15

[30] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996. 16

[31] P. Földiák, "Learning invariance from transformation sequences," *Neural Computation*, vol. 3, no. 2, pp. 194–200, 1991. 16, 20, 46, 48, 50, 56, 64, 82, 85

[32] G. Wallis and H. H. Bülthoff, "Effects of temporal association on recognition memory.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 4800–4, Apr. 2001. 17, 46, 54, 55, 57, 64, 85

[33] N. Li and J. J. DiCarlo, "Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex," *Neuron*, vol. 67, no. 6, pp. 1062–1075, 2010. 17, 24, 46, 53, 54, 55, 56, 64, 85

[34] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Magic materials: a theory of deep hierarchical architectures for learning sensory representations." 2013. 18, 27

[35] L. Wiskott and T. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002. 20, 46, 48, 64, 85

[36] T. Poggio, J. Mutch, F. Anselmi, J. Z. Leibo, L. Rosasco, and A. Tacchetti, "The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work)," *MIT-CSAIL-TR-2012-035*, 2012. 21, 67, 94, 101, 107

[37] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, (Lake Tahoe, CA), pp. 1106–1114, 2012. 21, 22

[38] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277–4280, 2012. 21, 22

[39] C. F. Cadieu, H. Hong, D. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo, "The Neural Representation Benchmark and its Evaluation on Brain and Machine," *arXiv preprint*, vol. arXiv:1301, 2013. 22

[40] T. Poggio, J. Mutch, F. Anselmi, A. Tacchetti, L. Rosasco, and J. Z. Leibo, "Does invariant recognition predict tuning of neurons in sensory cortex?," *MIT-CSAIL-TR-2013-019, CBCL-313*, 2013. 22, 24

[41] H. Bülthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Proceedings of the National Academy of Sciences*, vol. 89, no. 1, p. 60, 1992. 23, 62, 66, 75, 101

[42] M. J. Tarr and I. Gauthier, "Do viewpoint-dependent mechanisms generalize across members of a class?," *Cognition*, vol. 67, no. 1, pp. 73–110, 1998. 23, 101

[43] M. Dill and S. Edelman, "Imperfect invariance to object translation in the discrimination of complex shapes," *Perception*, vol. 30, no. 6, pp. 707–724, 2001. 25, 66

[44] I. Biederman and E. Cooper, "Evidence for complete translational and reflectional

invariance in visual object priming," *Perception*, vol. 20, no. 5, pp. 585–593, 1991. 25

[45] H. Op de Beeck and R. Vogels, "Spatial sensitivity of macaque inferior temporal neurons," *The Journal of Comparative Neurology*, vol. 426, no. 4, pp. 505–518, 2000. 25

[46] M. Dill and M. Fahle, "Limited translation invariance of human visual pattern recognition," *Perception and Psychophysics*, vol. 60, no. 1, pp. 65–81, 1998. 26, 34

[47] M. Dill and M. Fahle, "The role of visual field position in pattern-discrimination learning.," *Proceedings of the Royal Society B: Biological Sciences*, vol. 264, no. 1384, p. 1031, 1997. 26

[48] T. A. Nazir and K. J. O'Regan, "Some results on translation invariance in the human visual system," *Spatial Vision*, vol. 5, pp. 81–100, Jan. 1990. 26, 66

[49] J. Kahn and D. Foster, "Visual comparison of rotated and reflected random-dot patterns as a function of their positional symmetry and separation in the field," *The Quarterly Journal of Experimental Psychology Section A*, vol. 33, no. 2, pp. 155–166, 1981. 26, 34, 66

[50] V. H. Perry and A. Cowey, "The ganglion cell and cone distributions in the monkey's retina: implications for central magnification factors," *Vision Research*, vol. 25, no. 12, pp. 1795–810, 1985. 26

[51] G. Wallis and E. T. Rolls, "Invariant face and object recognition in the visual system," *Progress in Neurobiology*, vol. 51, pp. 167–194, 1997. 26, 46, 48

[52] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 15, pp. 6424–6429, 2007. 26, 37, 73, 81, 111

[53] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, “A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex,” *CBCL Paper #259/AI Memo #2005-036*, 2005. 26

[54] L. Isik, J. Z. Leibo, J. Mutch, S. W. Lee, and T. Poggio, “A hierarchical model of peripheral vision,” *MIT-CSAIL-TR-2011-031, CBCL-300*, 2011. 26

[55] T. Poggio, “The Computational Magic of the Ventral Stream: Towards a Theory,” *Nature Precedings*, July 2011. 27

[56] Q. Liao, J. Z. Leibo, and T. Poggio, “Learning invariant representations and applications to face verification,” in *Advances in Neural Information Processing Systems (NIPS)*, (Lake Tahoe, CA), 2013. 27, 83, 103

[57] J. Z. Leibo, Q. Liao, and T. Poggio, “Subtasks of Unconstrained Face Recognition,” in *Under Review*. 27

[58] D. Green and J. Swets, *Signal detection theory and psychophysics*. Los Altos, CA, USA: Peninsula Publishing, 1989. 28, 34, 65, 92

[59] R. Goris and H. Op De Beeck, “Neural representations that support invariant object recognition,” *Frontiers in Computational Neuroscience*, vol. 4, no. 12, 2010. 31

[60] D. Zoccolan, M. Kouh, T. Poggio, and J. J. DiCarlo, “Trade-off between object selectivity and tolerance in monkey inferotemporal cortex,” *Journal of Neuroscience*, vol. 27, no. 45, p. 12292, 2007. 31, 81

[61] N. Li, D. Cox, D. Zoccolan, and J. J. DiCarlo, “What response properties do individual neurons need to underlie position and clutter" invariant" object recognition?,” *Journal of Neurophysiology*, vol. 102, no. 1, p. 360, 2009. 32

[62] E. M. Meyers, D. Freedman, G. Kreiman, and T. Poggio, “Dynamic population coding of category information in inferior temporal and prefrontal cortex,” *Journal of neurophysiology*, vol. 100, no. 3, p. 1407, 2008. 33

[63] T. A. Carlson, H. Hogendoorn, R. Kanai, J. Mesik, and J. Turret, "High temporal resolution decoding of object position and category," *Journal of Vision*, vol. 11, no. 10, 2011. 33

[64] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in real-life images: Detection, alignment and recognition (ECCV)*, (Marseille, Fr), 2008. 34, 105, 106, 120

[65] H. Barrett, C. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions," *Journal of the Optical Society of America-A-Optics Image Science and Vision*, vol. 15, no. 6, pp. 1520–1535, 1998. 35

[66] N. Troje and H. Bülthoff, "Face recognition under varying poses: The role of texture and shape," *Vision Research*, vol. 36, no. 12, pp. 1761–1771, 1996. 40, 67, 87, 101, 107

[67] S. Ullman and S. Soloviev, "Computation of pattern invariance in brain-like structures," *Neural Networks*, vol. 12, pp. 1021–1036, Oct. 1999. 40

[68] F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004," in *Workshop on Generative-Model Based Vision*, vol. 2, 2004. 43

[69] R. M. Rifkin, *Everything old is new again: a fresh look at historical approaches in machine learning*. PhD thesis, Massachusetts Institute of Technology, 2002. 43

[70] N. Pinto and D. Cox, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," *IEEE Automated Face and Gesture Recognition (FG)*, vol. 25, pp. 26–27, 2011. 44

[71] W. Einhäuser, J. Hipp, and J. Eggert, “Learning viewpoint invariant object representations using a temporal coherence principle,” *Biological Cybernetics*, 2005. 46, 48

[72] M. Franzius, H. Sprekeler, and L. Wiskott, “Slowness and sparseness lead to place, head-direction, and spatial-view cells,” *PLoS Computational Biology*, vol. 3, no. 8, p. e166, 2007. 46, 48

[73] T. Masquelier and S. J. Thorpe, “Unsupervised learning of visual features through spike timing dependent plasticity,” *PLoS Computational Biology*, vol. 3, no. 2, 2007. 46, 48, 56

[74] T. Masquelier, T. Serre, S. Thorpe, and T. Poggio, “Learning complex cell invariance from natural videos: A plausibility proof,” *AI Technical Report #2007-060 CBCL Paper #269*, 2007. 46, 48, 64

[75] M. Spratling, “Learning viewpoint invariant perceptual representations from cluttered images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 753–761, 2005. 46, 48, 64

[76] R. Wyss, P. König, and P. Verschure, “A model of the ventral visual system based on temporal stability and local memory,” *PLoS biology*, vol. 4, no. 5, pp. 1–20, 2006. 46

[77] G. Wallis, B. Backus, M. Langer, G. Huebner, and H. Bülthoff, “Learning illumination-and orientation-invariant representations of objects through temporal association,” *Journal of vision*, vol. 9, no. 7, 2009. 46, 55, 57, 64

[78] I. Gauthier and M. J. Tarr, “Becoming a" greeble" expert: Exploring mechanisms for face recognition,” *Vision Research*, vol. 37, no. 12, pp. 1673–1682, 1997. 47

[79] T. Vetter, A. Hurlbert, and T. Poggio, “View-based models of 3D object recognition: invariance to imaging transformations,” *Cerebral Cortex*, vol. 5, no. 3, p. 261, 1995. 57, 63

[80] S. Ku, A. Tolias, N. Logothetis, and J. Goense, "fMRI of the Face-Processing Network in the Ventral Temporal Lobe of Awake and Anesthetized Macaques," *Neuron*, vol. 70, no. 2, pp. 352–362, 2011. 60, 76, 99

[81] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, no. 6676, pp. 598–601, 1998. 60, 100

[82] L. Cohen, S. Dehaene, and L. Naccache, "The visual word form area," *Brain*, vol. 123, no. 2, p. 291, 2000. 60, 76, 100

[83] C. Baker, J. Liu, L. Wald, K. Kwong, T. Benner, and N. Kanwisher, "Visual word processing and experiential origins of functional selectivity in human extrastriate cortex," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, p. 9087, 2007. 60

[84] P. Downing and Y. Jiang, "A cortical area selective for visual processing of the human body," *Science*, vol. 293, no. 5539, p. 2470, 2001. 60, 76

[85] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt and Co., Inc., 1982. 61

[86] S. Moeller, W. A. Freiwald, and D. Tsao, "Patches with links: a unified system for processing faces in the macaque temporal lobe," *Science*, vol. 320, no. 5881, p. 1355, 2008. 61

[87] J. DiCarlo, D. Zoccolan, and N. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012. 62, 107

[88] N. Logothetis, J. Pauls, H. Bülthoff, and T. Poggio, "View-dependent object recognition by monkeys," *Current Biology*, vol. 4, no. 5, pp. 401–414, 1994. 62, 66, 75, 79, 101

[89] M. J. Tarr and H. Bülthoff, "Image-based object recognition in man, monkey and machine," *Cognition*, vol. 67, no. 1, pp. 1–20, 1998. 62, 101

[90] T. Vetter and T. Poggio, “Linear object classes and image synthesis from a single example image,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 733–742, 2002. 63

[91] J. Mutch and D. Lowe, “Object class recognition and localization using sparse features with limited receptive fields,” *International Journal of Computer Vision*, vol. 80, no. 1, pp. 45–57, 2008. 64

[92] S. Stringer and E. Rolls, “Invariant object recognition in the visual system with novel views of 3D objects,” *Neural Computation*, vol. 14, no. 11, pp. 2585–2596, 2002. 64

[93] T. Poggio and S. Edelman, “A network that learns to recognize three-dimensional objects,” *Nature*, vol. 343, no. 6255, pp. 263–266, 1990. 64, 101

[94] D. Perrett and M. Oram, “Neurophysiology of shape processing,” *Image and Vision Computing*, vol. 11, no. 6, pp. 317–333, 1993. 64, 81

[95] B. W. Mel, “SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition,” *Neural Computation*, vol. 9, pp. 777–804, May 1997. 64

[96] N. Pinto, J. J. DiCarlo, and D. Cox, “How far can you get with a modern face recognition test set using only simple features?,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2591–2598, IEEE, 2009. 64, 111, 118

[97] L. Isik, J. Z. Leibo, and T. Poggio, “Learning and disrupting invariance in visual recognition with a temporal association rule,” *Front. Comput. Neurosci.*, vol. 6, no. 37, 2012. 64, 85

[98] P. Downing and M. Peelen, “The role of occipitotemporal body-selective regions in person perception,” *Cognitive Neuroscience*, vol. 2, no. 3-4, pp. 186–203, 2011. 73

[99] 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*, vol. 361, no. 1476, p. 2109, 2006. 76

[100] M. Peelen and P. Downing, "Selectivity for the human body in the fusiform gyrus," *Journal of Neurophysiology*, vol. 93, no. 1, pp. 603–608, 2005. 76

[101] R. Malach, J. B. Reppas, R. R. Benson, 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*, vol. 92, no. 18, pp. 8135–8139, 1995. 76, 99

[102] Blender.org, "Blender 2.6," 2013. 78, 108

[103] Singular Inversions, "FaceGen Modeller 3," 2003. 78, 108

[104] S. Nene, S. Nayar, and H. Murase, "Columbia Object Image Library (COIL-100)," *Columbia University Tech. Report No. CUCS-006-96*, 1996. 79

[105] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of mathematical biology*, vol. 15, no. 3, pp. 267–273, 1982. 82, 86, 94

[106] R. Douglas and K. Martin, "Neuronal circuits of the neocortex," *Annu. Rev. Neurosci.*, vol. 27, pp. 419–451, 2004. 84

[107] M. Sur, P. Garraghty, and A. Roe, "Experimentally induced visual projections into auditory thalamus and cortex," *Science*, vol. 242, no. 4884, p. 1437, 1988. 84

[108] Y. Miyashita, "Neuronal correlate of visual associative long-term memory in the primate temporal cortex," *Nature*, vol. 335, no. 6193, pp. 817–820, 1988. 85

[109] S. M. Stringer, G. Perry, E. T. Rolls, and J. Proske, "Learning invariant object recognition in the visual system with continuous transformations," *Biological cybernetics*, vol. 94, no. 2, pp. 128–142, 2006. 85

[110] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Wiley, 1949. 86, 94

[111] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature Reviews Neuroscience*, vol. 5, no. 2, pp. 97–107, 2004. 86, 94

[112] T. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural networks*, vol. 2, no. 6, pp. 459–473, 1989. 94

[113] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Networks*, vol. 5, no. 6, pp. 927–935, 1992. 94

[114] J. Mutch, J. Z. Leibo, S. Smale, L. Rosasco, and T. Poggio, "Neurons That Confuse Mirror-Symmetric Object Views," *MIT-CSAIL-TR-2010-062, CBCL-295*, 2010. 96

[115] S. Ramon y Cajal, *Texture of the Nervous System of Man and the Vertebrates: I*. Springer, 1999. 98

[116] H. Barlow, "Why have multiple cortical areas?," *Vision Research*, vol. 26, no. 1, pp. 81–90, 1986. 98

[117] G. Mitchison, "Neuronal branching patterns and the economy of cortical wiring," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 245, no. 1313, pp. 151–158, 1991. 98, 99

[118] D. B. Chklovskii and A. A. Koulakov, "Maps in the brain: What can we learn from them?," *Annual Review of Neuroscience*, vol. 27, pp. 369–392, 2004. 98

[119] E. Marder, "Neuromodulation of neuronal circuits: back to the future," *Neuron*, vol. 76, no. 1, pp. 1–11, 2012. 99

[120] A. Yu and P. Dayan, "Uncertainty, neuromodulation, and attention," *Neuron*, vol. 46, no. 4, pp. 681–692, 2005. 99

[121] E. Y. Ko, J. Z. Leibo, and T. Poggio, “A hierarchical model of perspective-invariant scene identification,” in *Society for Neuroscience (486.16/OO26).*, (Washington D.C.), 2011. 100

[122] M. Bar, E. Aminoff, and D. L. Schacter, “Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se,” *The Journal of Neuroscience*, vol. 28, no. 34, pp. 8539–8544, 2008. 100

[123] D. L. Schacter and D. 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*, vol. 364, no. 1521, pp. 1245–1253, 2009. 100

[124] D. Marr and H. K. Nishihara, “Representation and recognition of the spatial organization of three-dimensional shapes,” *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 200, no. 1140, pp. 269–294, 1978. 101

[125] I. Biederman, “Recognition-by-components: a theory of human image understanding.,” *Psychological review*, vol. 94, no. 2, p. 115, 1987. 101

[126] S. Ullman, “Aligning pictorial descriptions: An approach to object recognitionâĂć 1,” *Cognition*, vol. 32, no. 3, pp. 193–254, 1989. 101

[127] M. J. Tarr and H. H. Bülthoff, “Is human object recognition better described by geon structural descriptions or by multiple views?,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 6, pp. 1494–1505, 1995. 101

[128] P. G. Schyns, “Diagnostic recognition: task constraints, object information, and their interactions,” *Cognition*, vol. 67, no. 1, pp. 147–179, 1998. 101

[129] H. Hill, P. G. Schyns, and S. Akamatsu, “Information and viewpoint dependence in face recognition,” *Cognition*, vol. 62, no. 2, pp. 201–222, 1997. 101

[130] L. Lo Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri, "Spectral algorithms for supervised learning," *Neural Computation*, vol. 20, no. 7, pp. 1873–1897, 2008. 103

[131] N. Kanwisher, "Functional specificity in the human brain: a window into the functional architecture of the mind," *Proceedings of the National Academy of Sciences*, vol. 107, no. 25, p. 11163, 2010. 104

[132] A. Tarski, "What are logical notions?," *History and philosophy of logic*, vol. 7, no. 2, pp. 143–154, 1986. 104

[133] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 105, 106, 121

[134] N. Pinto, Y. Barhomi, D. Cox, and J. J. DiCarlo, "Comparing state-of-the-art visual features on invariant object recognition tasks," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pp. 463–470, IEEE, 2011. 105, 113, 114

[135] N. Pinto, D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?," *PLoS computational biology*, vol. 4, no. 1, p. e27, 2008. 106, 111

[136] P. Grother, G. Quinn, and P. Phillips, "Report on the evaluation of 2d still-image face recognition algorithms," *NIST Interagency Report*, vol. 7709, 2010. 106

[137] W. Braje, D. Kersten, M. Tarr, and N. Troje, "Illumination effects in face recognition," *Psychobiology*, vol. 26, no. 4, pp. 371–380, 1998. 107

[138] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005. 110

[139] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Ma-*

chine Intelligence, IEEE Transactions on, vol. 32, no. 9, pp. 1627–1645, 2010. 110

[140] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008. 110

[141] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002. 110

[142] M. Guillaumin, J. Verbeek, and C. Chmid, "Is that you? Metric learning approaches for face identification," in *IEEE International Conference on Computer Vision*, (Kyoto, Japan), pp. 498–505, 2009. 110

[143] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1978–1990, 2011. 110, 119

[144] S. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. British Machine Vision Conference (BMCV)*, vol. 1, (Guildford, UK), pp. 52–61, 2012. 110

[145] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, pp. 236–243, Springer, 2008. 110

[146] C. Chan, M. Tahir, J. Kittler, and M. Pietikainen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1164–1177, 2013. 110

[147] J. Mutch, U. Knoblich, and T. Poggio, "CNS: a GPU-based framework for simulating cortically-organized networks," *MIT-CSAIL-TR*, vol. 2010-013, no. 286, 2010. 111

[148] N. Pinto, J. J. DiCarlo, D. D. Cox, *et al.*, "Establishing good benchmarks and baselines for face recognition," in *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008. 111, 118

[149] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999. 111

[150] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Empowering visual categorization with the gpu," *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp. 60–70, 2011. 112

[151] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004. 115

[152] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, (Providence, RI), pp. 2879–2886, 2012. 117

[153] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, vol. 1, pp. 947–954, IEEE, 2005.

[154] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.