Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2010-061
CBCL-294

December 30, 2010

Learning Generic Invariances in Object Recognition: Translation and Scale

Joel Z Leibo, Jim Mutch, Lorenzo Rosasco, Shimon Ullman, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Learning Generic Invariances in Object Recognition: Translation and Scale

Joel Z Leibo[1,2,3], Jim Mutch[1,2,3], Lorenzo Rosasco[1,2,3], Shimon Ullman[4], and Tomaso Poggio[1,2,3]

[1] *Center for Biological and Computational Learning, Cambridge MA USA*
[2] *McGovern Institute for Brain Research, Cambridge MA USA*
[3] *Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge MA USA*
[4] *Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot Israel*

Abstract

Invariance to various transformations is key to object recognition but existing definitions of invariance are somewhat confusing while discussions of invariance are often confused. In this report, we provide an operational definition of invariance by formally defining perceptual tasks as classification problems. The definition should be appropriate for physiology, psychophysics and computational modeling.

For any specific object, invariance can be trivially "learned" by memorizing a sufficient number of example images of the transformed object. While our formal definition of invariance also covers such cases, this report focuses instead on invariance from very few images and mostly on invariances from one example. Image-plane invariances – such as translation, rotation and scaling – can be computed from a single image for any object. They are called generic since in principle they can be hardwired or learned (during development) for any object.

In this perspective, we characterize the invariance range of a class of feedforward architectures for visual recognition that mimic the hierarchical organization of the ventral stream. We show that this class of models achieves essentially perfect translation and scaling invariance for novel images. In this architecture a new image is represented in terms of weights of "templates" (e.g. "centers" or "basis functions") at each level in the hierarchy. Such a representation inherits the invariance of each template, which is implemented through replication of the corresponding "simple" units across positions or scales and their "association" in a "complex" unit. We show simulations on real images that characterize the type and number of templates needed to support the invariant recognition of novel objects. We find that 1) the templates need not be visually similar to the target objects and that 2) a very small number of them is sufficient for good recognition.

These somewhat surprising empirical results have intriguing implications for the learning of invariant recognition during the development of a biological organism, such as a human baby. In particular, we conjecture that invariance to translation and scale may be learned by the association – through temporal contiguity – of a small number of primal templates, that is patches extracted from the images of an object moving on the retina across positions and scales. The number of templates can later be augmented by bootstrapping mechanisms using the correspondence provided by the primal templates – without the need of temporal contiguity.

This version replaces a preliminary CBCL paper which was cited as: Leibo et al. "Invariant Recognition of Objects by Vision," CBCL-291, November 2, 2010

Part I

Invariance: definition

1 Introduction

How is it that we can recognize objects despite the extreme variability of the retinal images that they may generate? Consider translation invariance: an object translating across an organism's visual field activates an entirely different set of photoreceptors when it is on the left versus when it is on the right. Somehow the visual system must associate all the various patterns evoked by each object so that perceptual invariance is maintained at higher levels of processing.

The retinotopy of the neural representation of visual space persists as information is passed from the retina to the brain. The receptive fields of neurons in primary visual cortex inherit the spatial organization of the retina. In higher visual areas cells respond to increasingly large regions of space (Desimone and Schein, 1987; Logothetis et al., 1995). At the end of this processing hierarchy, in the most anterior parts of the ventral visual system (in particular: AIT) there are cells that respond invariantly despite significant shifts (up to several degrees of visual angle[1]).

Perhaps because of the strong interest in the problem of invariance from multiple scientific communities including neurophysiology, psychophysics and computer vision, multiple mutually inconsistent definitions coexist. In the first part of this report, we provide an operational definition of "invariance" – to translation and other transformations – that can be used across physiological and psychophysical experiments as well as computational simulations.

2 Defining invariance

2.1 Neural decoding and the measurement of population invariance

Invariance in neuronal responses can be studied at the single cell or population level. Even when "single-cell invariance" is nominally the property of interest, it is appropriate to think in terms of a population. In the case of a single-cell study, this means focusing on the population of inputs to the recorded cells, since a cell's responses can be regarded as the outputs of a classifer operating on its inputs. We interpret electrophysiology studies of neuronal populations in terms of the information represented in the cortical area being recorded from, while we interpret single-cell studies in terms of the information represented in the inputs to the recorded cells[2].

[1] The claim that IT neurons respond invariantly to shifts in stimulus position has had a long and somewhat twisted history since Gross et al. first recorded single-unit activity from that region in the 1960s. Initial reports described receptive fields always larger than 10 by 10 degrees, some fields more than 30 by 30 degrees and one cell responding everywhere on the 70 by 70 degree screen (Gross et al., 1969). Since these early reports, the trend has been toward more careful measurement and smaller average field size estimates. However, even the smallest of the resulting measurements still show a substantial proportion of cells in IT cortex with larger receptive fields than are present in striate cortex (DiCarlo and Maunsell, 2003) and the claim that visual representations increase in position tolerance as they traverse the ventral stream from V1 towards IT is not in doubt.

[2] In the case of a population consisting of a single cell, a specialization of our proposed definition of AuT (see section 2.4 reduces to "rank-order invariance" as proposed by Logothetis et al. (1995) and Li et al. (2009).

Whether we consider a single cell or a population, the critical issue is whether or not the representation enables invariant readout. All physiological measures of invariance involve the adoption of a particular method of decoding information from neural responses. The primary difference between apparently distinct measures is often the choice of decoding method. This becomes a particularly thorny issue when measuring population invariance. The visual task being modeled plays an important role in the choice of decoding method. Most decoding is done using a classifier with parameters that are typically fit using training data (Hung et al., 2005; Li et al., 2009; Liu et al., 2009; Meyers et al., 2008; Quiroga et al., 2007). In this conceptual framework, in order to address questions about invariant recognition of novel objects encountered for the first time at a particular location or scale, the classifier must be trained using a single training example and tested with an appropriate universe of distractors. In contrast, the easier task of invariantly recognizing familiar objects with prior experience at many locations is modeled with a classifier trained using more examples. Our operational definitions encompass both these situations and allow the direct comparison of results from both types of experiments as well as many others.

2.2 The problem of invariance

We are mostly concerned with the problem of recognizing a target object when it is presented again at a later time, perhaps after undergoing a transformation. In our framework, an image is generated by an object. The image is then measured by a brain or a seeing machine. It is on the basis of these measurements that a decision as to the identity of the object must be reached.

In our specific problem, an image of a *target* object has already been presented and measured. 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*.

2.3 The classification point of view

We formulate an operational definition of invariance range in terms of classification accuracy. We will use standard measures of classification performance and remark about trade-offs between these measures. Some of these performance measures are related to selectivity while others are related to invariance. We proceed by first defining a quantity called *Accuracy under Transformation* (AuT). We then define the *Invariance Range* in terms of AuT. The following section of this report (2.4) contains a more formal version of this same definition.

We start with the specialization of the definition for the case of translation invariance for novel objects which is easy to generalize to other transformations and object classes.

Definition *Choose a disk of radius r and a universe U of target and distractor objects. Train a classifier C with the target at the center of the disk. 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. We define Accuracy under Transformation as a summary statistic describing the classifier's performance on this test and indicate it as $AuT_{C,U}(r)$.*

Remark: We can analogously define $AuT_{C,U}$ for any transformation. For example, we could use rotated or scaled versions of target and distractor images in order to define scaling or rotation invariance.

Remark: *AuT* is defined in general as a function of a set of parameters. This may be the radius of the region over which objects could be translated in the case of translation invariance or a maximum allowable rotation angle in the case of rotation invariance on the plane.

2.4 Formal development of the classification point of view

Let X denote the set of all images of targets and distractors. For a test image $x \in X$ there are two possibilities:

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

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

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

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

We can now define some measures of the classifier's performance on this task.

$$\begin{aligned} TP(\eta) &:= \int_X P(y = 1, C(x) \geq \eta | x) p(x) dx = \text{True positive rate} \\ FP(\eta) &:= \int_X P(y = -1, C(x) \geq \eta | x) p(x) dx = \text{False positive rate} \\ TN(\eta) &:= \int_X P(y = -1, C(x) < \eta | x) p(x) dx = \text{True negative rate} \\ FN(\eta) &:= \int_X P(y = 1, C(x) < \eta | x) p(x) dx = \text{False negative rate} \end{aligned}$$

Note: The probability $p(x)$ of picking any particular image is normally assumed to be uniform.

Remark: 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 a machine learning classifer operating on data.

Example: In psychophysics and signal detection theory the underlying distributions of target $P_P = P(x, y = 1)$ and distractor $P_N = P(x, y = -1)$ images are assumed to be Gaussian over the familiarity $C(x)$. Any choice of classifier and decision criterion apportions the probability mass of P_P and P_N over both X_A^η and X_R^η giving us the situation depicted in figure 1.

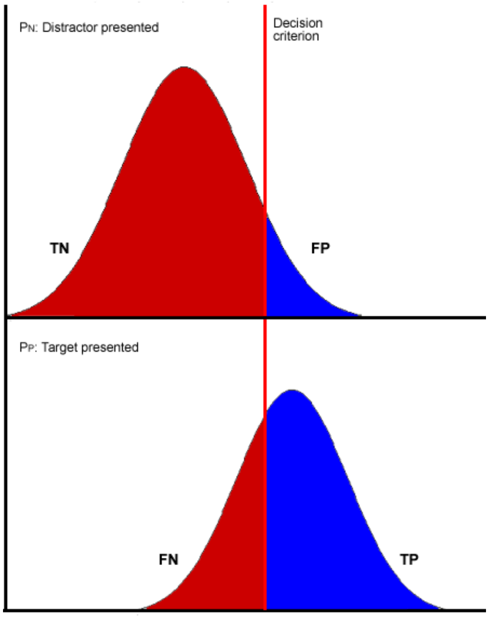

Figure 1: *Example distributions in the case where P_P and P_N are Gaussian over $C(x)$. The top panel shows an example distribution P_N of test images arising in the case that $y = -1$ and the bottom panel shows an example distribution P_P of test images in the case the $y = 1$. The performance measures: TP (true positive), TN (true negative), FP (false positive) and FN (false negative) rate correspond to the quantities obtained by integrating each distribution over the regions shown here.*

The goal of this decision task is to accept test images judged to contain the same object as the target image and reject images judged not to depict the target. In this case "invariance" is related to the rate of acceptances, i.e. judging more test images to be the same as the target, while "selectivity" is related to the opposite tendency to reject - judging more test images to be different from the target image. For any decision criterion we get a picture like the one shown in figure 1. In this picture, the blue region is related to invariance while the red region is related to selectivity. We will refer to the acceptance rate (blue region) and the rejection rate (red region).

Remark: *For η a decision criterion:*

$$\text{Acceptance rate} \propto TP(\eta) + FP(\eta) \tag{1}$$

$$\text{Rejection rate} \propto TN(\eta) + FN(\eta) \tag{2}$$

If we change the decision criterion to optimize any performance measure e.g. $TP(\eta), FP(\eta)$, $TN(\eta), FN(\eta)$ then there must be a concomitant decrease in some other performance measures. For example, increasing $TP(\eta)$ -a good thing- must be accompanied by an increase in $FP(\eta)$ -a bad thing.

The measure AuT - defined informally above and formally below - is a *bias-free* summary statistic. It provides information about the *tradeoff* of acceptance rate - related to invariance, and rejection rate - related to selectivity. See section 2.5 for additional discussion of this point.

2.4.1 Defining Accuracy under Transformation (AuT)

The performance measures considered so far depend on η. Varying η generates the operating characteristic (ROC) curve[3]:

$$ROC(\eta) = [FP(\eta), TP(\eta)]\, \forall \eta \tag{3}$$

Note: All values of $ROC(\eta)$ will fall on the unit square, $(0 \leq TP(\eta) \leq 1)$ and $(0 \leq FP(\eta) \leq 1)$, because both quantities are integrals of probability distributions. 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 (AuT).

Definition: Accuracy under Transformation *For X a set of images with labels y and P_P the distribution of targets on X, P_N the distribution of distractors on X, 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 under Transformation $AuT_{C,X}$ is the area under the ROC curve:*

$$AuT_{C,X} = \int_0^1 \overline{ROC}(z)dz \tag{4}$$

Remark: It is simple to extend this operational definition to study parametrized transformations such as translation, scaling and rotation.

Let X be the union of a sequence of sets of images X_i ordered by inclusion.

$$X = \bigcup_i X_i \text{ with } X_i \subset X_{i+1} \forall i$$

Then we can compute the corresponding AuT_{C,X_i} for each index i.

As an example, you could study translation invariance by letting X_i contain all the images of target and distractor objects at each position in a circle of radius r_i. Subsequent sets X_{i+1} contain objects at each position within radius $r_{i+1} > r_i$ thus $X_i \subset X_{i+1}$.

Remark: In discussions of invariance we often want to speak about relevant and irrelevant dimensions of stimulus variation (Goris and Op De Beeck, 2010; Zoccolan et al., 2007). For example, the shape of the object could be the relevant dimension while its position in space may be the irrelevant dimension. Our notion of Accuracy under Transformation extends to capture this situation as well. To illustrate, we consider the case of two parametrized dimensions. Let i, j index the union $X = \bigcup_{i,j} X_{i,j}$. So for each pair i, j there is a corresponding subset of X. We require that:

$$X_{i,j} \subset X_{i+1,j} \text{ and } X_{i,j} \subset X_{i,j+1}\ \forall (i,j)$$

Accuracy under Transformation may be computed for each pair of indices i, j using equation 4. If one of the stimulus dimensions is not continuous then this situation is easily accomodated by letting either i or j take on only a few discrete values.

[3] In the case that $C(x)$ is a likelihood ratio test, the Neyman-Pearson lemma guarantees that the ROC curve will be concave and monotonic-increasing. Insofar as all other classifiers approximate likelihood ratio tests, then they too will usually induce concave and monotonic-increasing ROC curves.

2.5 Selectivity and invariance

It is not possible to give operational definitions for "selectivity" and "invariance" that encompass all the common usages of these terms. The interesting cases that may be called selectivity-invariance trade-offs arise when Accuracy under Transformation (AuT) declines when considering increasingly extreme transformations (e.g. translating over larger regions). If $AuT(r) = AuT$ does not depend on the translation range r then classification performance can be said to be "invariant" to translation.

In the context of a same-different task there is another sense of the selectivity-invariance trade-off. The classifier's acceptance rate is related to invariance while its rejection rate is related to selectivity. This captures the intuition that invariance is a tendency to accept most inputs and selectivity is a tendency to reject all but a few inputs.

In summary, when discussing the "selectivity-invariance trade-off" it is important to be clear. If selectivity and invariance are associated with the classifier's rejection and acceptance rates then there is a trivial trade-off that can be seen by varying the decision criterion (the bias). On the other hand, there are situations in which there is a more interesting trade-off between Accuracy under Transformation and r – the size of the region over which targets and distractors can appear. These two trade-offs have very different interpretations: the former can be attributed to bias differences between classifiers while the latter reflects more fundamental aspects of the transformation or the representation of the stimuli. Care must be taken in experimental work not to conflate these two trade-offs.

2.6 The invariance range

We can define the *invariance range* by picking a threshold level of Accuracy under Transformation θ and determining the maximal region size r for which $AuT(r)$ 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 under Transformation for each.

Definition: Invariance range *Let X_i be a sequence of sets of images ordered by inclusion, C_i an ordered sequence of classifiers and $AuT(i) = AuT_{C_i,X_i}$ be the classification accuracy obtained by using $C_i(x)$ to partition X_i. Let θ be a threshold value. Then the invariance range I is the maximal i for which $AuT(i) > \theta$.*

$$I = \begin{cases} \infty & AuT(i) > \theta \quad \forall i \\ \max\{i | AuT(i) > \theta\} & \text{otherwise} \end{cases} \tag{5}$$

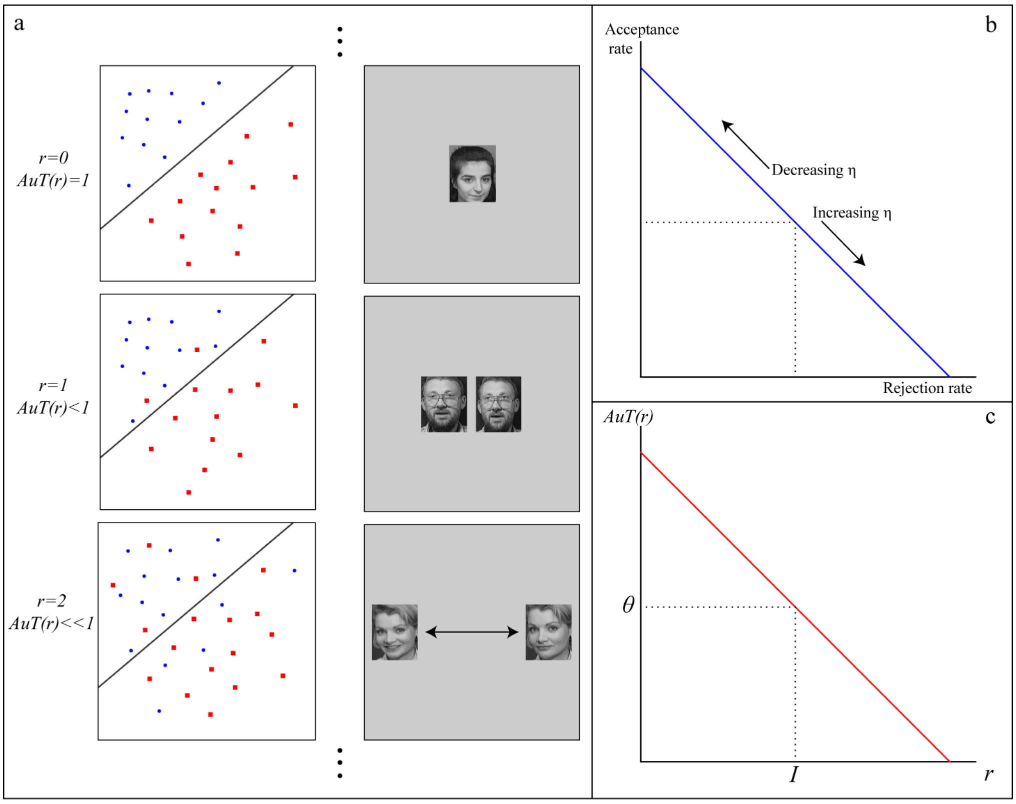

Figure 2: *a) Example of separating hyperplanes in a feature space (left) for variable distances r over which targets and distractors may appear. Possible most extreme translations of the target are shown in the right column corresponding to the AuT on the left. b) For a particular value of r (range over which objects may appear), there is a trivial trade-off between acceptance rate and rejection rate obtained by varying the decision criterion η. c) Another trade-off may occur across values of r. The invariance range I is the maximal r for which $AuT(r)$ is above a threshold θ.*

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 $AuT(r)$ 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.

3 Measuring Accuracy under Transformation

3.1 Physiology

We can measure $AuT(r)$ 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. $AuT(r)$ 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 various other publications (Hung et al., 2005; Meyers et al., 2008).

Remark: We can also measure AuT 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.

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 AuT 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. Dill and Fahle (1998); Kahn and Foster (1981), 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.

3.3 Psychophysics

We can compute AuT 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 (Green and Swets, 1989). In this case AuT is related to the standard psychophysical measure of discriminability d' by the following:

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

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 Barrett et al. (1998) for a simple derivation of this relationship.

Part II

Invariance in hierarchical models

4 A hierarchical model of invariant object recognition

4.1 Motivation from physiology

Extracellular recordings from cat V1 in the early 1960s by David Hubel and Torsten Wiesel yielded the observation of simple cells responding to edges of a particular orientation. Hubel and Wiesel also described another class of cells with more complicated responses which came to be called complex cells. In the same publication (Hubel and Wiesel, 1962), they hypothesized that (at least some of) the complex cells may be receiving their input from the simple cells.

The simple cells' receptive fields contain oriented "on" regions in which presenting an edge-stimulus excited the cell and "off" regions for which stimulus presentation suppressed firing. These classical "Gabor-like" receptive fields can be understood by noting that they are easily built from a convergence of inputs from the center-surround receptive fields of the lateral geniculate nucleus (LGN). The V1 simple cells respond selectively when receiving an input from several LGN cells with receptive fields arranged along a line of the appropriate orientation. Figure 3A is a reproduction of Hubel and Wiesel's original drawing from their 1962 publication illustrating the appropriate convergence of LGN inputs.

In contrast to simple cells, Hubel and Wiesel's complex cells respond to edges with particular orientations but notably have no off regions where stimulus presentation reduces responses. Most complex cells also have larger receptive fields than simple cells, i.e. an edge of the appropriate orientation will stimulate the cell when presented anywhere over a larger region of space. Hubel and Wiesel noted that the complex cell fields could be explained by a convergence of inputs from simple cells. Figure 3B reproduces their scheme.

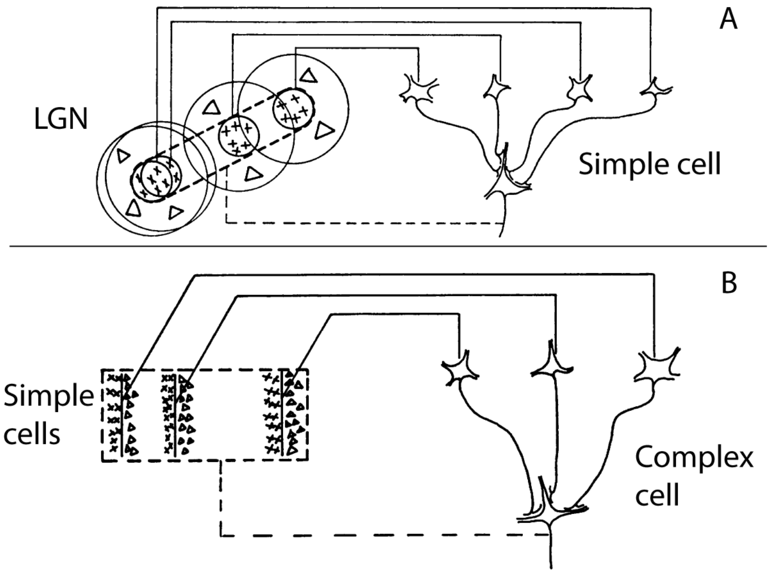

Figure 3: *Adapted from (Hubel and Wiesel, 1962).*

Following Hubel and Wiesel, we say that the simple cells are tuned to a particular preferred feature. This tuning is accomplished by weighting the LGN inputs in such a way that a simple cell fires when the inputs arranged to build the preferred feature are co-activated. In contrast, the complex cells' inputs are weighted such that the activation of any of their inputs can drive the cell by itself. So the complex cells are said to pool the response of several simple cells. As a visual signal passes from LGN to V1 its representation increases in selectivity; patterns without edges (such as sufficiently small circular dots of light) are no longer represented. Then as the signal passes from simple cells to complex cells the representation gains in invariance. Complex cells downstream from simple cells that respond only when their preferred feature appears in a small window of space now represent stimuli presented over a larger region. Notice also that simple cells could be implemented on the dendritic tree of complex cells (Serre et al., 2005).

4.2 Model implementation

At the end of the hierarchy of visual processing, the cells in IT respond selectively to highly complex stimuli and also invariantly over several degrees of visual angle. A popular class of models of visual processing subject an input signal to a series of selectivity-increasing and invariance-increasing operations (Fukushima, 1980; Perrett and Oram, 1993; Riesenhuber and Poggio, 1999). Higher level representations become tuned to more and more complex preferred features through selectivity-increasing operations and come to tolerate more severe identity-preserving transformations through invariance-increasing operations.

We implemented such a biologically-plausible model of the visual system modified from (Serre et al., 2007a). This 4-layer model converts images into a feature representation via a series of processing stages referred to as layers. In order, the layers of the model were: S1 → C1 → S2 → C2. In our model, an object presented at a position A will evoke a particular pattern of activity in layer S2. When the object is moved to a new position B, the pattern of activity in layer S2 will change accordingly. However, this translation will leave the pattern in the C2 layer unaffected.

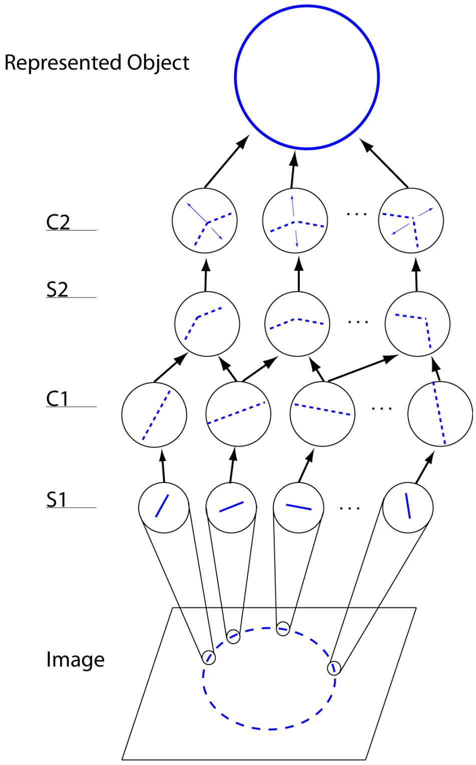

Figure 4: *An illustration of the hierarchical model of object recognition.*

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. The S1 units model the response of Hubel and Wiesel's V1 simple cells. 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. We regard the C1 units as modeling Hubel and Wiesel's V1 complex cells. Each layer labeled S is computing a selectivity-increasing operation while the C layers perform invariance-increasing operations.

The S2 units employ a template-matching operation to detect features of intermediate complexity. The preferred features of S2 units are preprocessed versions of small patches extracted from natural images. Here "preprocessed" means that the template-matching operation is performed on the output of the previous layer and so is encoded as the pattern of activity of a set of C1 units. The S2 units compute the following function of their inputs $x = (x_1, ...x_n)$

$$r = \exp\left(-\frac{1}{2\sigma}\sum_{j=1}^{n}(w_j - x_j)^2\right) \tag{6}$$

where the unit's preferred feature is encoded in the stored weights $w = (w_1, ..., w_n)$ and σ is a parameter controlling the tightness of tuning to the preferred feature. A large σ value would make the response tolerate large deviations from its preferred feature while a small sigma value will cause the unit to respond only when the input closely resembles its preference.

Following Serre et al. (2007a) we chose the preferred features of S2 units by randomly sampling patches from a set of natural images and storing them (C1-encoded) as the S2 weights. So the response of an S2 unit to a new image can be thought of as the similarity of the input to a previously encountered template image. An S2 representation of a new object is a vector of these similarities to previously-acquired templates.

A practically equivalent and biologically more plausible operation at the level of S units involves a normalized dot product instead of the operation described in equation 6.

4.3 Invariance simulations

First we consider a model in which we have replicated each S2 unit at nearly every position in the visual field. In all our simulations C2 units pool over the entire visual field, receiving input from all S2 units with a given template. Thus at the top level of the model, there will be exactly one C2 unit for each template.

4.3.1 Simulation methods

Unless stated otherwise, all of the following 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. That is, for translation simulations the trained image depicted the target at the center of the visual field, for scaling simulations the trained image showed the target at the intermediate size and for rotation simulations the trained image depicted the object at 0 degrees (straight ahead).

All the AUC (equivalently: AuT) 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.

4.3.2 Simulation results

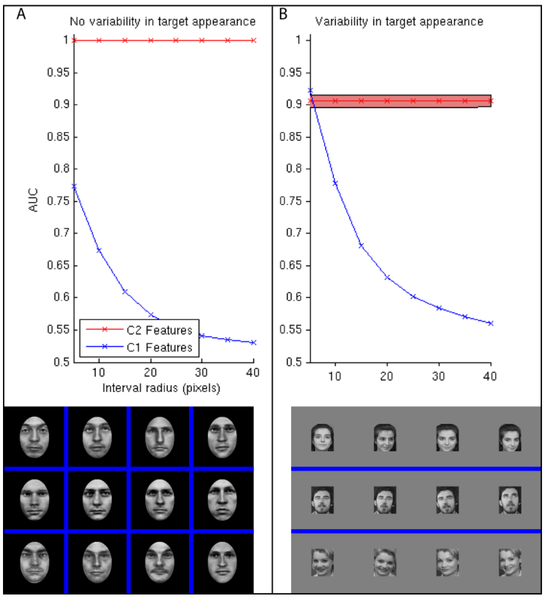

Figure 5: *Accuracy under Translation over intervals of increasing size. The size of the interval over which targets and distractors could appear is plotted as the abscissa with the corresponding AuT as the ordinate. For these simulations there were 2000 C2 layer cells with patches randomly sampled from 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 (Troje and Bülthoff, 1996).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.*

The model performs perfectly on the simple face discrimination task shown in the left panel of figure 5 (red curve). In this version of the model, each S2 unit is replicated at every position in the visual field. C2 units corresponding to each template pool over all the corresponding S2 units at every position. This C2 representation is sufficiently selective and invariant that perfect performance is obtained. That is, the representation of a single target image is always more similar to itself across translation than it is to any other image; the invariance range is effectively infinite. When running the classifier on the C1 representation (blue curve) we obtain a different result: these units have much smaller pooling ranges and thus do not yield position-invariant performance. That is, $AuT(r)$ declines with increasing distance from the trained location and invariance range is finite.

In figure 5 panel B, the task was made slightly more difficult. Now the positive class contains multiple images of each person under small variations in pose and expression. The task is to rank all the images in the positive class as being more similar to one another than to any of the distractor images. From the single training example we used (one example face presented in the center of the visual field) the resulting Accuracy under Translation is imperfect (red curve). However, invariance range is unaffected by this change in task. i.e. *AuT* does not depend on the radius of the interval over which targets and distractors could appear.

The invariance range for scaling transformations is also nearly infinite. Accuracy under Scaling was imperfect due to discretization effects in the scale pyramids of the model (see figure 6). The decline of *AuT* with scaling was modest and attributable to discretization effects.

The results in figure 7 appear to show infinite invariance range for translation over clutter. However, there are some caveats to this interpretation. The cluttered backgrounds used here are not very similar to the to-be-recognized objects. Clutter is a problem for object recognition when it is - in the words of Stuart Geman - *"made up of highly structured pieces that will conspire to mimic an object"* (Geman, 2006). The cluttered backgrounds used in this simulation are probably not sufficiently similar to the test objects for this to occur.

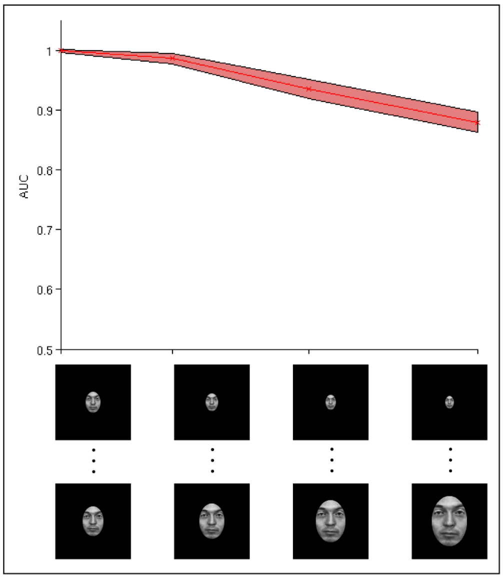

Figure 6: *Scale invariance. This model included scale pyramids at each level. The S1 layer contained 10 scales which utilized the same scaling factors as the test images and each C1 unit pooled over two adjacent scales. The C2 layer pooled over all scales. Templates for S2 / C2 were generated from 2000 patches of natural images. See Serre et al. (2007b) and Mutch and Lowe (2008) for details on these scale pyramids. The classifier used only the vector of C2 responses evoked by each image. Error bars shown here are +/-1 standard deviation across runs with different templates. The abscissa shows the range of scales over which test images appeared. This test of scale invariance is analogous to the one in figure 5A for translation invariance i.e. there was no variability in the appearance of the target face except that which was produced by scaling.*

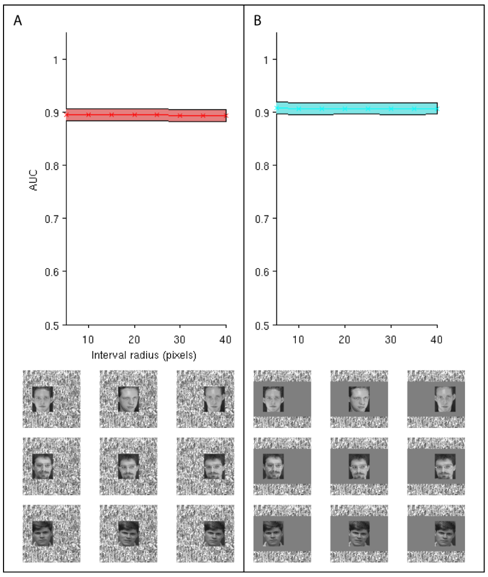

Figure 7: *The effect of clutter. Model and testing procedure was the same as in Figure 5B. Training was always done with uncluttered images: our classifier measured the C2-layer similarity of each cluttered test image to a single uncluttered image presented in the center. Error bars show the standard deviation over runs with different sets of S2/C2 templates generated by randomly sampling patches from natural images. Panel A: Full clutter condition. Clutter filled the background of the image. As the object translated, it uncovered different parts of the background clutter. Panel B: Partial clutter condition. Clutter did not fill the image; the object never occluded any part of the cluttered background.*

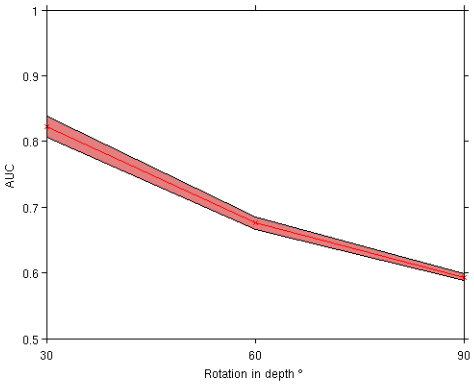

Figure 8: *3D-rotation-in-depth: 10 testing faces for identity recognition were each rendered at 7 viewpoints using specialized 3D face modeling software (Facegen- Singular Inversions Inc.). None of the faces had hair or other features of high rotation-invariant diagnostic value. All faces had the same skin color. This simulation was conducted using the "full" version of the model which contains features at 10 different scales. This is similar to the model used for the simulation shown in figure 6. The abscissa denotes the range of rotation away from* 0° *(straight-ahead view) over which faces were presented. Faces could rotate in either direction. Templates for S2/C2 layers were generated from 2000 patches of natural images. The classifier used only the vector of C2-layer responses evoked by each testing image. Error bars show +/-1 standard deviation across runs with different templates.*

4.3.3 Implications for physiology

The results in figures 5, 6 and 7 found invariance ranges for translation and scaling that are, in principle, infinite. In contrast, a single-unit electrophysiology study in macaque AIT by Zoccolan et al. (2007) found a significant negative relationship between the proportion of images that evoke responses from a cell (sparseness) and that cell's tolerance to translation and scaling transformations. That is, cells with high sparseness also tended to have low tolerance and highly tolerant cells also tended to respond less sparsely. They interpret these results to mean that there is a trade-off between selectivity (sparseness) and invariance (tolerance). On the other hand, our modeling results seem to suggest that, at least for translation and scaling, selectivity need not be lost in order to gain invariance.

As mentioned in section 2.5, care must be taken in interpreting experimental results as evidence of a fundamental selectivity-invariance trade-off. The Zoccolan et al. (2007) results could be accounted for in two ways:

1. AIT cells have different biases, that is, they are all operating in different regimes of their ROC curve. This is the "trivial" selectivity-invariance trade-off.
2. The population of inputs to AIT cells employ a representation that is not similar to that used by the C2 layer of our model. For example, the inputs to AIT cells could be more similar to the C1 layer of our model in which case a trade-off between Accuracy under Transformation and tolerance is expected (see figure 5).

If these results are attributable to bias differences between cells then it remains possible that the inputs to AIT employ a C2-like representation. It is also possible that AIT represents an intermediate processing stage, perhaps like our S2 layer, and invariant representations arise at a later stage of processing. Note: a computational simulation showed that this pattern of results can be obtained for S2-like cells in our model by varying the number of afferents or the tuning width parameter σ (Zoccolan et al., 2007). These two accounts of the electrophysiology results are not mutually exclusive and indeed both may occur.

5 Invariance for novel objects

5.1 The role of hierarchy

In order to achieve invariant recognition the neural pattern evoked by the object must be associated with the patterns evoked by transformed versions of the same object. Given a collection of patterns evoked by the same object presented in every position, various unsupervised and supervised methods exist to learn which to associate in order to achieve invariance (see section 5.3). It is tempting to conclude that invariant recognition of novel objects should be impossible since no examples of the transformed object are available for training. We have referred to this problem as *the puzzle of initial invariance*. However, it is really a non-problem for hierarchical models of object recognition. Novel objects can be encoded by their similarity to a set of templates (loosely, we can think of these as object parts or better as patches or fragments extracted from images of the object). In these models, invariance for novel objects can be inherited from the invariant detection of templates. We can interpret the templates as familiar object parts that have previously been encountered, and associated with each other, over a variety of transformations.

In the situation originally described by Hubel and Wiesel, there is a simple cell with each preferred orientation replicated at every position in the visual field. The complex cells are thus able to gain in invariance by pooling over several simple cells at a range of locations with the same preferred orientation. If complex cells accurately pick out cells with the appropriate preferred orientation as well as their translated counterparts, then the complex cell will come to respond selectively to edges of a particular orientation presented at a range of spatial locations. This scheme can be repeated at higher levels of the hierarchy, thus generating invariance for more complicated template patterns.

Pooling in general, and max-pooling in particular, provides an idealized mathematical description of the operation obtained after accurately associating templates across transformations. In the brain, these associations must be learned[4]. Moreover, hierarchical models of object recognition typically work by first creating local feature detectors at every position and then pooling over all their responses to create an invariant template. This approach is only possible in the case of image-plane transformations, e.g. position, scale and 2D in-plane rotation. In these cases it is possible to compute, from a single example, the appearance of the template under the full range of transformation parameters. We refer to such transformations for which this approach is available as *generic* and note that there are also non-generic transformations such as viewpoint and illumination changes. The appearance of an object under non-generic transformations depends on its 3D structure or material properties and thus, in these cases, it is not possible to obtain invariance by creating transformed features and pooling.

In all the simulations described in the previous section, input images were compared to the pattern evoked by the target object presented at the center of the receptive field. This method of classifying input images by the similarity of their evoked responses to those evoked by the target object involves no training on distractors or transformed versions of the target. Therefore, the classifier models the task of invariantly recognizing a novel object. Figures 5, 6 and 7 show that this particular biologically-plausible model of object recognition achieves translation and scale invariant performance for novel faces. Figure 8 shows that, for novel faces, the model is not invariant to 3D rotation in depth, a non-generic transformation.

5.2 Psychophysics of initial invariance

We have shown that hierarchical models can recognize novel objects invariantly of position and scale, but can humans do this? A number of psychophysical studies have directly investigated human performance on these tasks.

For psychophysical experiments with human observers, it is not trivial to pick a class of novel objects for which the observers will truly be unfamiliar. Dill and Edelman (2001) showed that human 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. However, it may be that the stimuli in that experiment were not sufficiently novel. Similar experiments using random dot pattern stimuli showed that discriminability declines when objects are presented as little as 2 degrees away from the trained location (Dill and Fahle, 1997, 1998; Kahn and Foster, 1981; Nazir and O'Regan, 1990). However, despite drops in performance with translation, accuracy remains significantly above chance. Notably, Dill and Fahle (1998) showed that recognition performance when objects appear at a novel location is the same as the performance obtained before training at the to-be-trained location and significantly above chance. They speculate that

[4] Perhaps by Hebbian mechanisms operating over time. See section 5.3 of this report.

there could be two recognition processes at work in these experiments with different properties and developing over different time scales. A translation invariant mechanism accounts for the initial performance with one or very few training examples while a slow-to-develop non-invariant mechanism accounts for improved performance with increasing training at a particular location. We interpret our computational simulations of translation invariance as providing a biologically-plausible account of the initial invariance observed in these experiments. It is likely that the slow-to-develop position-dependent mechanism is related to other perceptual learning effects in the literature (Poggio et al., 1992) which are not addressed by this model. The slow-to-develop effects could be accounted for by memorization of object-specific, localized templates.

5.3 Learning invariance

Replicating each template under all transformation parameters (e.g. positions) and pooling over all the resulting local feature detectors is a method to obtain invariance for generic transformations. This method can be regarded as modeling the final outcome of an associative learning process that associates the neural patterns evoked by an object seen under transformations. It is important to consider the learning algorithms that could bring about this outcome.

Many biologically plausible candidate learning algorithms are motivated by the temporal statistics of the visual world. 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 of natural vision 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 templates (Földiák, 1991; Wiskott and Sejnowski, 2002).

There are numerous proof-of-principle computational systems that implement variations on temporal association algorithms to learn invariant representations from sequences of images (Franzius and Wilbert, 2008; Oram and Foldiak, 1996; Spratling, 2005; Stringer and Rolls, 2002; Wallis and Rolls, 1997) and natural videos (Kayser et al., 2001; Masquelier et al., 2007). Psychophysical studies test 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 (Wallis and Bülthoff, 2001). 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 occured (Cox et al., 2005).

There is also physiological evidence that the brain utilizes temporal associations in order to learn invariant templates. Li and DiCarlo (2008) showed that exposure to an altered visual environment in which highly-dissimilar objects swap identity across saccades causes AIT neurons to change their stimulus preference at the swapped location. They went on to obtain similar results for scale invariance; 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[5] (Li and DiCarlo, 2010).

The biological mechanism underlying the brain's implementation of learning by temporal association remains unknown. However, Spike-Timing Dependent Plasticity (STDP) (Bi and Poo, 1998;

[5] Notably, a control unswapped location and scale was unaffected in both experiments (Li and DiCarlo, 2008, 2010)

Markram et al., 1997) is a popular proposal for which computational simulations establishing the plausiblity of the approach have been carried out (Masquelier and Thorpe, 2007).

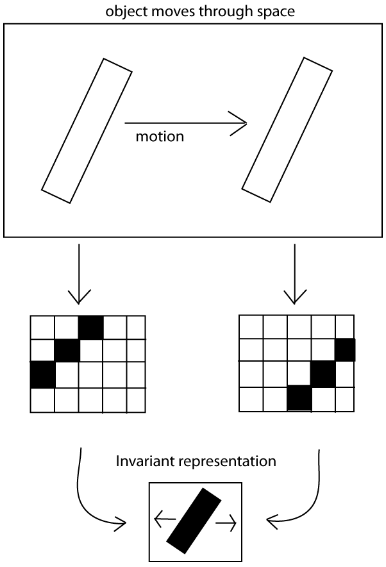

Figure 9: *Illustration of how invariance could be learned from a temporal sequence of images depicting an object translating across the visual field.*

5.4 Do templates need to be learned from similar objects?

Prior visual experience allows the neural patterns evoked by a template under different transformations to be accurately associated, thereby enabling invariant detection of templates. We have shown that invariant recognition of novel objects is made possible by encoding novel objects as their similarity to a set of familiar invariant templates. In this account, invariance for novel objects is inherited from learned invariance for familiar component parts (templates).

It is a folk wisdom that templates only enable recognition of novel objects that resemble the templates or are built from them (as for instance Ullman's fragments are (Ullman, 1999)). We show that the alternative hypothesis is true: any object can be encoded by its similarity to essentially any set of templates.

As evident from figure 10 (bottom right panel), even invariant templates extracted from highly unnatural objects (random dot patterns) are sufficient to support invariant face identification. There are two[6] sources of variability in the target and distractor images used for this task: the variability due to translating the stimuli and the variability induced by the nature of the task (multiple appearances for each target). The former is resolved by the use of invariant templates.

Learning invariant templates from images similar to those in the test set produces a representation that better reflects the most diagnostic aspects of the test images and thus affects overall accuracy

[6] Actually there is a third source of variability in the patterns induced by the target. This variability is due to the discretization period of the recognition system. We do not consider this variability to be an unbiological artifact of our model's architecture. Biologically implemented feature detectors also experience the exact same retinal position mismatches that give rise to the discretization effect in our model. In our model implementation (and likely in biology as well), this effect is small compared to the variability introduced by the task itself.

level (Serre et al., 2007b). However, invariance range is independent of overall accuracy. A set of invariant feature detectors that lead to the target often being confused with a distractor will continue to do so at all eccentricities.

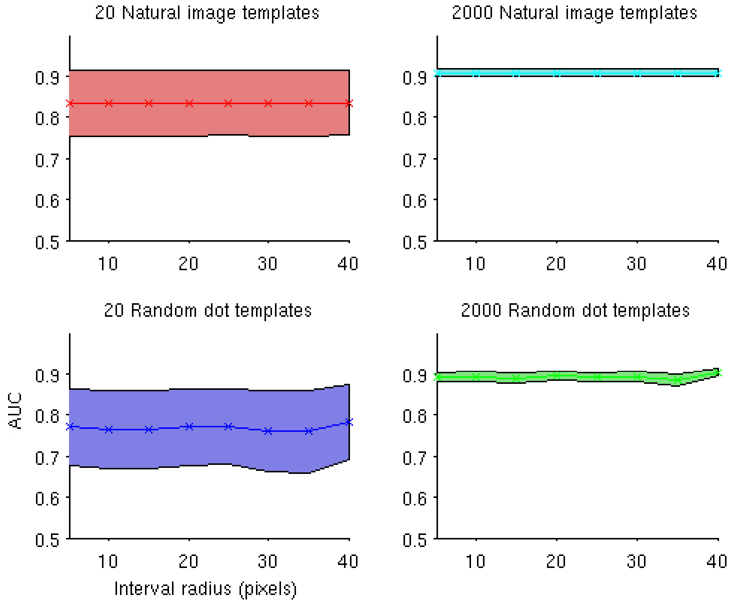

Figure 10: *AUC (AuT) for recognizing targets versus distractors appearing anywhere within an interval of a particular radius (using C2-layer features only). Red curve: 20 translation invariant features were extracted from natural images. Cyan curve: 2000 invariant features extracted from natural images. Blue and green curves: 20 and 2000 translation invariant features extracted from random dot patterns respectively. We build invariant templates 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 5B.*

5.5 How many templates are necessary?

After establishing that recognition does not in itself require the detection of features matched to the test objects we can now attempt to discover the minimal requirements for invariant recognition. Figure 10 (left panels) displays the results of two simulations we ran utilizing only a very small number of invariant templates (20) drawn from images of either random dot patterns (blue curve) or natural images (red curve). Surprisingly, these results show that invariance range is maintained despite using a very small number of templates.

When small numbers of templates are employed, accuracy is more affected by the similarity of the templates to the particular test images. That is, if you only have a few invariant templates it is helpful to have extracted them from similar objects to the test set. Figure 11 shows classification accuracy as a function of the number and type of invariant feature detectors employed.

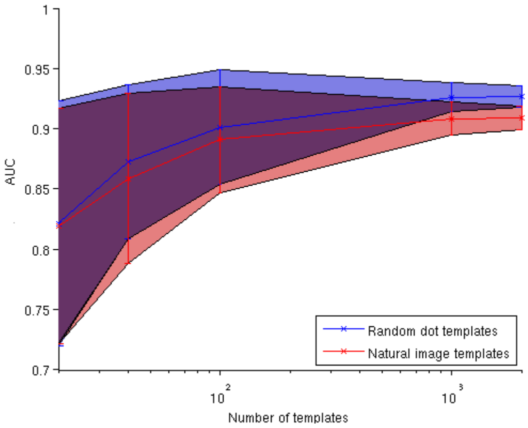

Figure 11: *AUC (AuT) for the face identification task as a function of the number of templates used to build the model. Here invariant templates were derived from random patches of 20 natural images (red curve) or 20 random dot patterns (blue curve). Restarting the simulation and choosing different 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 5B.*

The test images used for the above simulations contained some variability of shape in the positive set of target images. The classifier was tasked with ranking all the test images by their similarity to the image of a particular face presented in the center of the visual field. Each face was included in several versions with slight variations in pose, expression and lighting. We also tested the same model on a dataset including even more severe variability within the positive class: the Caltech 101 set of images (Li et al., 2004); see figure 12.

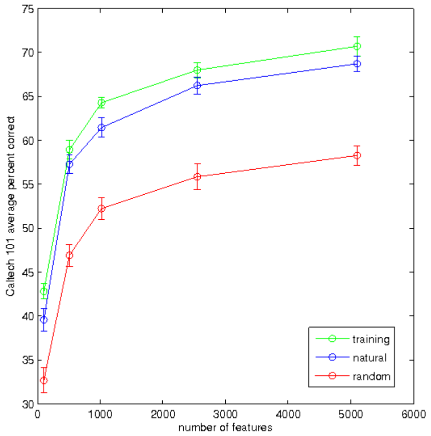

Figure 12: *Average percent correct on the Caltech 101 dataset using a regularized least squares (RLS) classifier with 30 training images per class. The templates were extracted from random patches drawn from either the training images (green curve), an unrelated set of natural images (blue curve) or 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.*

These Caltech 101 tests revealed that templates drawn from random dot patterns yield considerably worse performance than templates drawn from natural images. However, templates drawn from random dots still support a level of accuracy that is far above chance[7]. Accuracy on this more difficult task increases when we increase the number of templates employed.

5.6 Summary

Invariant recognition of novel objects is possible by encoding novel objects in terms of their similarity to a set of invariant templates. Furthermore, it is possible to learn a set of invariant templates by temporal association of templates obtained from an object undergoing transformations during natural vision. We showed that this method of learning invariance for the recognition of any novel objects works for generic image transformations such as translation and scaling.

We determined that the invariant templates need not be similar to the novel to-be-recognized objects in order to support invariant recognition. Furthermore, only a very small number of templates are needed to ensure invariance over the entire receptive field. Classification accuracy, however, does increase with the use of greater numbers of templates and more task-suitable templates.

In the final section of this report we discuss an extension of these ideas to motivate new learning algorithms that can acquire invariant templates without the requirement of temporal association.

[7] Chance would be less than 1% correct on the Caltech101 set of images.

Part III

Bootstrapping invariant object recognition

Upon first eye opening, a human infant is bombarded by visual stimuli. An early developmental task is to organize this input into objects and learn to recognize them despite identity-preserving transformations. In the previous sections we have provided computational evidence that a newborn infant could acquire a relatively small set of invariant templates through the association of convenient transforming features in its environment. We call these templates the *primal templates*[8].

Our computer experiments show good performance using a small number of invariant templates that do not even resemble the to-be-recognized objects. Thus a very small number of objects – possibly just one – moving across and within the visual field may enable the learning of a sufficient set of primal templates allowing invariant recognition. However, due to their small numbers and lack of similarity to the newly-encountered objects, many objects may be confused and accuracy will not be high enough.

Utilizing the primal templates, the infant's visual system could add new invariant templates. The invariant response of the primal templates could be used to establish correspondence between patches of images of the same object after undergoing an identity-preserving transformation. Importantly, this method would not require temporal continuity between the patches. With such a bootstrapping mechanism, the brain could bootstrap from an initially small set of primal templates into a much more complex system capable of representing the full richness of the visual world.

[8] In this view, primal templates may come from any objects moving in the infant's environment. For instance, patches of a mother's face may serve this role.

References

Barrett, H., Abbey, C., and Clarkson, E. (1998). 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*, 15(6):1520–1535.

Bi, G. and Poo, M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464.

Cox, D., Meier, P., Oertelt, N., and DiCarlo, J. (2005). 'Breaking'position-invariant object recognition. *Nature Neuroscience*, 8(9):1145–1147.

Desimone, R. and Schein, S. (1987). Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835.

DiCarlo, J. and Maunsell, J. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology*, 89(6):3264.

Dill, M. and Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception*, 30(6):707–724.

Dill, M. and Fahle, M. (1997). The role of visual field position in pattern-discrimination learning. *Proceedings of the Royal Society B: Biological Sciences*, 264(1384):1031.

Dill, M. and Fahle, M. (1998). Limited translation invariance of human visual pattern recognition. *Perception and Psychophysics*, 60(1):65–81.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200.

Franzius, M. and Wilbert, N. (2008). Invariant object recognition with slow feature analysis. *Artificial Neural Networks-ICANN 2008*, pages 961–970.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Geman, S. (2006). Invariance and selectivity in the ventral visual pathway. *Journal of Physiology-Paris*, 100(4):212–224.

Goris, R. and Op De Beeck, H. (2010). Neural representations that support invariant object recognition. *Frontiers in Computational Neuroscience*, 4(12).

Green, D. and Swets, J. (1989). *Signal detection theory and psychophysics.* Peninsula Publishing, Los Altos, CA, USA.

Gross, C., Bender, D., and Rocha-Miranda, C. (1969). Visual Receptive Fields of Neurons in Inferotemporal Cortex of Monkey. *Science*, 166:1303–1306.

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

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866.

Kahn, J. and Foster, D. (1981). 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*, 33(2):155–166.

Kayser, C., Einhäuser, W., and Dümmer, O. (2001). Extracting slow subspaces from natural videos leads to complex cells. *Artificial Neural Networks - ICANN*, pages 1075–1080.

Li, F.-F., Fergus, R., and Perona, P. (2004). 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*, volume 2.

Li, N., Cox, D., Zoccolan, D., and DiCarlo, J. (2009). What response properties do individual neurons need to underlie position and clutter" invariant" object recognition? *Journal of Neurophysiology*, 102(1):360.

Li, N. and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–7.

Li, N. and DiCarlo, J. J. (2010). Unsupervised Natural Visual Experience Rapidly Reshapes Size-Invariant Object Representation in Inferior Temporal Cortex. *Neuron*, 67(6):1062–1075.

Liu, H., Agam, Y., Madsen, J., and Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2):281–290.

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

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297):213.

Masquelier, T., Serre, T., Thorpe, S., and Poggio, T. (2007). Learning complex cell invariance from natural videos: A plausibility proof. *AI Technical Report*, #2007-069.

Masquelier, T. and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology*, 3(2).

Meyers, E., Freedman, D., and Kreiman, G. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology*, 100(3):1407.

Mutch, J. and Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57.

Nazir, T. A. and O'Regan, K. J. (1990). Some results on translation invariance in the human visual system. *Spatial Vision*, 5(2):81–100.

Oram, M. and Foldiak, P. (1996). Learning generalisation and localisation: Competition for stimulus type and receptive field. *Neurocomputing*, 11(2-4):297–321.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333.

Poggio, T., Fahle, M., and Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, 256(5059):1018–1021.

Quiroga, R., Reddy, L., Koch, C., and Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of neurophysiology*, 98(4):1997.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *CBCL Paper #259/AI Memo #2005-036.*

Serre, T., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–6429.

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

Spratling, M. (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):753–761.

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

Troje, N. and Bülthoff, H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771.

Ullman, S, S. S. (1999). Computation of pattern invariance in brain-like structures. *Neural Networks*, 12(7-8):1021–1036.

Wallis, G. and Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4800–4.

Wallis, G. and Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology*, 51:167–194.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27(45):12292.