

Invariant Recognition of Objects by Vision

Joel Z Leibo^{1,2,3}, Jim Mutch^{1,2,3}, Lorenzo Rosasco^{1,2,3}, Shimon Ullman⁴,
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. Image-plane invariances – such as translation, rotation and scaling – can be computed independently of the specific object. On the other hand, both invariance to rotation in depth and invariance to changes in illumination require implicit information about the 3D structure of the object or its material properties and thus more than a single “training” image. Here, we interpret same-different perceptual tasks as classification problems. This perspective allows us to provide a formal definition of the efficiency of invariance, a bias-free summary measure of the trade-off between selectivity and invariance. We believe that this definition is the most natural and should be used in physiology, psychophysics and modeling.

We characterized the efficiency of invariance in 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 perfect translation and scaling invariance for novel images. In this architecture a new image is represented in terms of weights of “templates” or “basis functions” at each level in the hierarchy. Such a representation inherits the invariance of the templates, which is built in through replications of the corresponding units across positions or scales. Simulations on real images characterize the type and number of templates needed for a representation which is sufficient to support the invariant recognition of novel objects.

We conclude that the templates need not be visually similar to the test objects and that using a very small number of them is sufficient for good recognition. This surprising empirical result yields intriguing implications for the learning of invariant recognition during the development of a biological organism, such as a human baby.

1 Introduction

How is it that we can recognize objects despite the extreme variability of the retinal impressions they may evoke? 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 have learned to associate all the various patterns evoked by each object so that perceptual invariance is maintained at a higher level of processing. In this report we will be concerned primarily with invariance to retinal position. However, most of our argument will apply more broadly to other identity-preserving transformations.

This separation of the representation of different parts 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. It is only as you consider higher-level visual areas such as V4 and IT that you begin to find cells that respond to larger 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 large shifts of several degrees of visual angle.

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 larger 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 ¹.

2 Defining invariance

2.1 The problem of invariance

We are 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 evoked by an object. The image is then measured in various ways by a brain or intelligent 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*.

In this situation we have the intuition that *invariance* manifests as a tendency to respond positively i.e. that the target and test images are identical. While *selectivity* is related to the tendency to respond negatively- that the target and test images are different.

2.2 The classification point of view

We describe this same-different recognition problem as a classification problem. We will define measures of classification performance and show that there are trade-offs between these measures. Some of these performance measures are related to selectivity while others are related to invariance.

¹Position tolerance must increase as information passes from V1 to IT. Consider that the smallest V1 receptive fields are found in the area representing the fovea. In that area V1 fields are often smaller than a degree with the smallest observed by Hubel and Wiesel to be around 10-15 arc minutes in size (Hubel and Wiesel, 1962). In IT, most receptive fields overlap the fovea, if we assume uncontroversially that IT cells receive visual inputs originating from the V1 cells representing the same region of space then it is clear that position tolerance must be increasing as the hierarchy is traversed from sub-degree field sizes in V1 to multi-degree fields in IT over the fovea.

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 a classifier $C : X \rightarrow \mathbb{R}$ 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 classifier 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 apports 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.

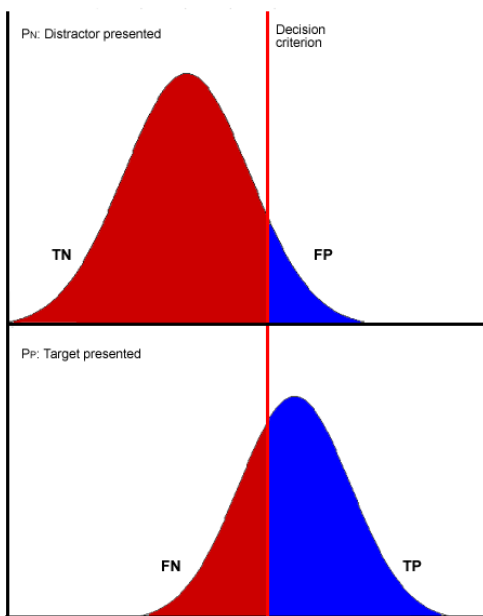


Figure 1: Example distributions in the case where P_P and P_N are Gaussians over $C(x)$. The top panel shows an example distribution P_N of test images arising in the case that $y = -1$ and the bottom panel shows an example distribution P_P of test images in the case the $y = 1$. The performance measures: 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 and x a test image then:

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

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

2.3 Thresholds and response bias

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.

We would like to utilize a bias-free summary statistic that provides information about the tradeoff of acceptance rate- related to invariance, and rejection rate- related to selectivity. Some problems may be shown to be “easy” in the sense that invariance can be gained without giving up selectivity and vice versa. While other problems may be more “difficult”; in which case, quite a lot of

selectivity must be given up to gain invariance. The measure of *efficiency of invariance* that we propose here is just such a measure.

2.4 Efficiency of invariance

All the performance measures we have considered so far are dependent on the decision criterion η . If we vary the criterion then we can draw out an operating characteristic (ROC) curve²:

$$ROC(\eta) = [FP(\eta), TP(\eta)] \forall \eta \quad (3)$$

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.

We propose to use the area under the ROC curve as an operational definition for the efficiency of invariance.

Definition: Efficiency of invariance 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 efficiency of invariance I is the area under the ROC curve:

$$I = \int_0^1 ROC(z) dz \quad (4)$$

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 efficiency of invariance 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 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 efficiency of invariance 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:

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

The efficiency of invariance 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 acomodated by letting either i or j take on only a few discrete values.

²In 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.

3 Measuring invariance

3.1 Physiology

We can measure the efficiency of invariance in physiology experiments. A typical experiment consists of an animal viewing stimuli, either passively or while engaged in a same/different 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 rate of a collection of cells in response to a stimulus. We can measure the efficiency of invariance in a population of any size i.e. this is neither a single-cell nor a population-only measure.

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. The efficiency of invariance is the area under the ROC curve.

Remark: It is important to choose a classifier that can be trained on the response to just a single image. The classifier must be trained only on the response to the target image presented in its original position.

Remark: A typical electrophysiology experiment 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 the efficiency of invariance 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 Computation

Most computer object recognition systems can be divided into two components. 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 vision-questions about the images e.g. are two test images displaying the same object? We can directly compute the efficiency of invariance using the classifier and the ground truth labels y_x .

Remark: Regarding the elements of the feature vector evoked by an image analogously to neuronal firing rates or fMRI voxels then the procedure for measuring the efficiency of invariance from computational experiments and physiological experiments is exactly the same.

3.3 Psychophysics

We can compute the efficiency of invariance in behavioral experiments. However, in this case we need to make 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 P_N and P_P are both Gaussian. This is the standard assumption of signal detection theory (Green and Swets, 1989). In this case the efficiency of invariance is related to the standard psychophysical measure of discriminability d' by the following:

$$I = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left(\frac{d'}{2} \right)$$

Where $\operatorname{erf}()$ denotes the error function:

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

and $d' = Z(TP) - Z(FP)$ where $Z()$ denotes a Z-score. See Barrett et al. (1998) for a simple derivation of this relationship.

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 contained 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 2A 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 2B reproduces their scheme.

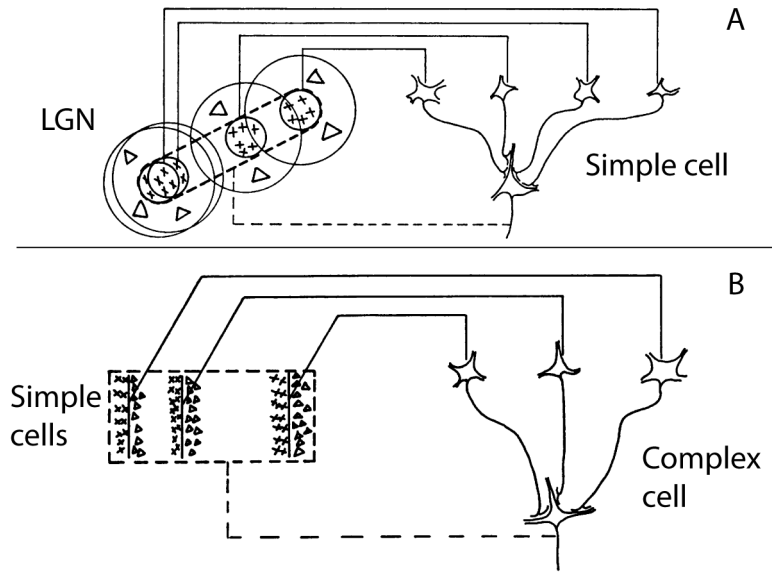


Figure 2: Adapted from (Hubel and Wiesel, 1962).

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.

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 proceed through subjecting 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 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: $S1 \rightarrow C1 \rightarrow S2 \rightarrow C2$. 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.

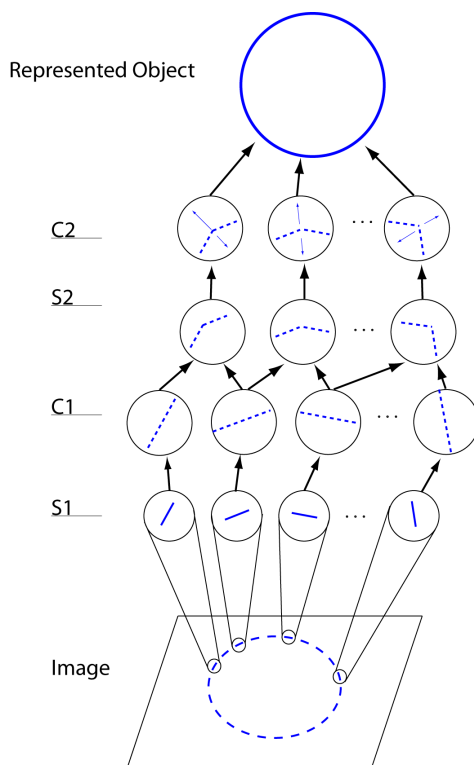


Figure 3: An illustration of the hierarchical model of object recognition.

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 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 in the pattern of activity of a set of C1 units. The S2 units compute the following function of their inputs $x = (x_1, \dots, x_n)$.

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

Where the unit's preferred feature is encoded in the stored weights $w = (w_1, \dots, w_n)$ 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) in 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.

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 stored.

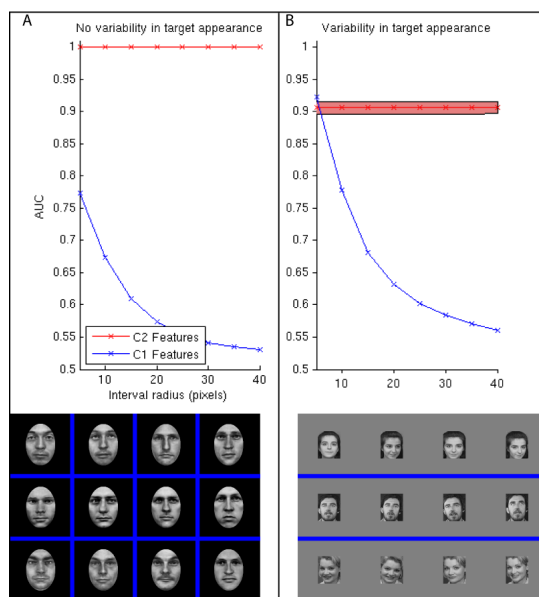


Figure 4: The efficiency of invariance over intervals of increasing size. The size of the interval over which targets and distractors could appear is plotted as the abscissa with the corresponding efficiency of invariance (AUC) as the ordinate. For these simulations there were 2000 C2 layer cells with patches randomly sampled from natural images. Panel A: The classifier computes the distance from the representation of a target face presented at the center to the representation of an input face presented at a variable location. The targets and distractors are faces modified from the Max Planck Institute face database (Troje and Bühlhoff, 1996). The images are 256x256 pixels and the faces are 120 pixels across. Panel B: The classifier still computes the distance 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 a different templates.

The model performs perfectly on the simple face discrimination task shown in the left panel of figure 4 (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. When running the classifier on the C1 unit representation (blue curve) we obtain a different result, these units with much smaller invariance ranges do not yield position-invariant performance. In this case the efficiency of invariance declines with increasing distance from the trained location.

In figure 4 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 efficiency is imperfect (red curve). However, the independence of efficiency invariance and interval radius is unaffected by this change in task. i.e. the AUC does not depend on the radius of the interval over which targets and distractors could appear.

The efficiency of invariance for scaling is also nearly independent of the range of the amount of scaling. Performance was imperfect due to discretization effects in the scale pyramids (see figure 5).

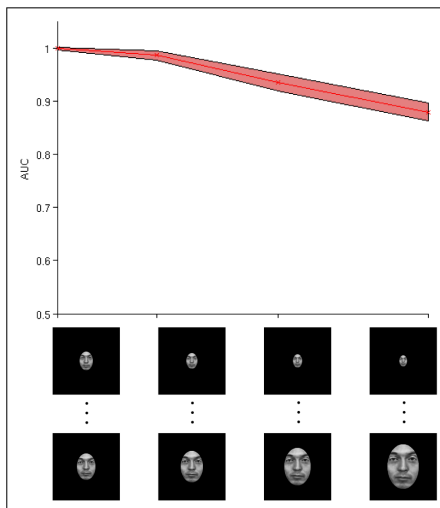


Figure 5: Scale invariance. This model included scale pyramids at each level. The S1 layer contained 10 scales which utilized the same scaling factors as the test images and each C1 unit pooled over two adjacent scales. The C2 layer pooled over all scales. Templates for S2 / C2 were generated from 2000 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 4A for translation invariance i.e. there was no variability in the appearance of the target face except that which was produced by scaling.

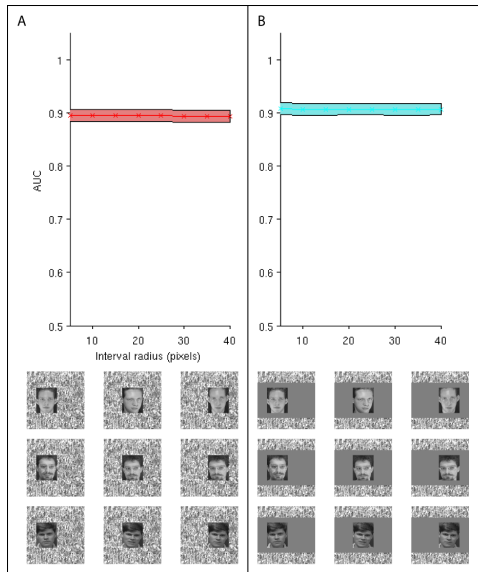


Figure 6: The effect of clutter. Model and testing procedure was the same as in Figure 4B. Training was always done with uncluttered images: our classifier measured the C2-layer similarity of each cluttered test image to a single uncluttered image presented in the center. Error bars show the standard 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.

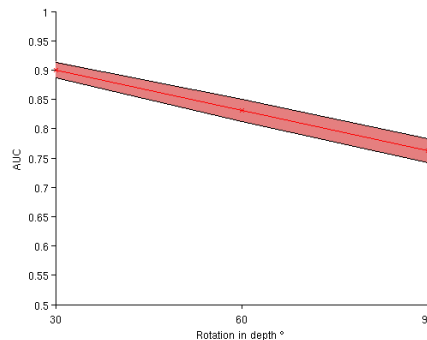


Figure 7: 3D rotation in depth. Testing images came from the Max Planck Institute for Biological Cybernetics face database (Troje and Bühlhoff, 1996). The abscissa denotes the range of rotation away from 0° (straight-ahead view) over which faces were presented. Faces could rotate in either direction. 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

4.4 The selectivity-invariance trade-off

There are two candidate phenomena that could go by the name of “selectivity-invariance trade-offs”. In the classification point of view (same/different task) that we have been considering, the

acceptance response rate, tendency to respond "same", is related to invariance and the rejection rate, tendency to respond "different", is related to selectivity. Changing the response criterion affects both response rates. Choosing a loose criterion has the effect of increasing the acceptance response rate and decreasing the rejection rate. Likewise, choosing a strict criterion has the opposite effect.

The second candidate phenomena for the name "selectivity-invariance tradeoff" is based on the efficiency of invariance. In some situations, the efficiency of invariance may vary with the range over the irrelevant stimulus dimension. Interpreting the terminology in DiCarlo and Cox (2007), we propose that their notion of *tangling* corresponds to this situation in which efficiency of invariance decreases when the range over which the irrelevant stimulus dimension is tested increases.

Remark: The results in figures 4,5 and 7 show that in our model, translation and scale are not tangled with object identity while 3D rotation in depth is tangled with identity.

4.4.1 Implications for physiology

Zoccolan et al. (2007) recorded from neurons in anterior inferotemporal cortex while a macaque monkey observed a library of images. This set of images included various transformed versions (translation, scale, contrast, etc). They found that some cells exhibited a high degree of sparsity (responding to very few images the library) and some cells were invariant over various transformations. However, there was a significant negative relationship between sparsity and their invariance measures for each transformation. Cells responding sparsely over the whole library of images, or over a subset where only shape was varied, were the same cells that were less invariant to all of the tested transformations. Likewise, cells that responded invariantly to any transformation were less likely to respond sparsely. We cannot yet determine whether the trade-off observed in this experiment is due to merely finding cells with variable decision criteria or if the brain employs a representation in which these stimulus dimensions are truly tangled.

5 Invariance for novel objects

5.1 The memory-based approach

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. In this case there need not be a concomitant decrease in selectivity. 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 method of obtaining invariant object recognition is often referred to as the memory-based approach. In the ideal situation, every feature detector of interest would be replicated under every transformation over which invariance is desired. Pooling over the appropriate replicated features yields the desired invariant detector. Whenever the dimensions over which selectivity and invariance are sought do not interact with one another there is no selectivity-invariance trade-off.

It is important to note that by referring to the memory-based approach we do not mean a strict grandmother cell look-up table scheme. It is possible to maintain the most attractive features of such a scheme while avoiding the combinatorial explosions (and other problems) with which look-up tables are plagued. This is done by encoding certain examples and interpolating between them. Various other papers from our group have discussed our view on this topic at length (Poggio, 1990; Poggio and Bizzi, 2004; Poggio and Edelman, 1990).

The memory-based approach to invariant object recognition is available whenever we can associate the patterns evoked by an object before and after the transformation for which we seek invariance. An object appearing in the left visual field will activate a completely different set of photoreceptors when presented on the right. The association of these highly disparate patterns by a higher-level unit is the essence of the memory-based approach to invariant recognition.

How then is the memory-based approach available at all? If invariant recognition requires associating highly dissimilar patterns evoked by the object under different conditions, then doesn't the system need to know which patterns were evoked by each object in order to associate them in the first place? Is there not some circular "if you've already solved the problem, solve it again that way" kind of logic to the memory-based approach?

Luckily, this difficulty with the memory-based approach can be overcome. The visual system is able to learn invariant representations without a priori knowing which patterns to associate. 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. The system could take advantage of this property of natural vision and adopt a rule that associates temporally contiguous patterns of activity. As such a system gains experience in the visual world it would gradually acquire invariant object representations (Földiák, 1991; Masquelier et al., 2007; Stringer and Rolls, 2002; Wiskott and Sejnowski, 2002).

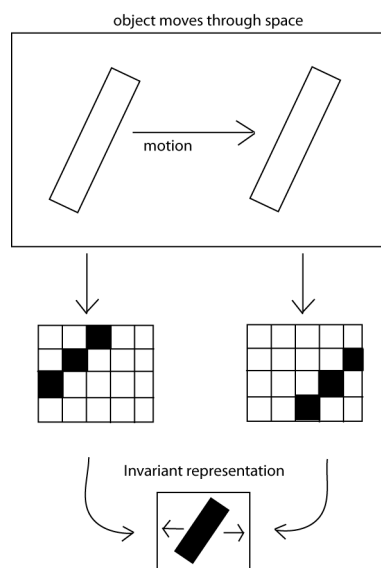


Figure 8: Illustration of the how invariance could be learned from a a temporal sequence of images depicting an object translating across the visual field.

5.2 The puzzle of initial invariance

Is prior experience with the objects to be recognized necessary? How can the memory-based approach explain our ability to recognize novel objects despite translation?. If an astronaut landed on Mars, emerged from her shuttle and began looking around at all the unfamiliar objects. It seems unlikely that she would be unable to recognize the new objects across movements of her head or eyes. How then is it possible for a memory-based approach that relies on previous viewings to predict invariance in this case? We refer to this problem as *the puzzle of initial invariance* and believe that its successful resolution is a crucial hurdle that all memory-based accounts of invariant-object recognition need to overcome.

5.3 Psychophysics and physiology of initial invariance and learning

Psychophysical evidence concurs with our intuitions that there is initial invariance even for novel objects. Dill and Edelman (2001) showed that human subjects could distinguish novel animal-like stimuli well above chance despite shifts away from the training location of (up to) 8 degrees of visual angle. Similar experiments also showed initial invariance significantly above chance for unfamiliar random dot pattern stimuli (Dill and Fahle, 1997, 1998; Nazir and O'Regan, 1990)³. Other psychophysical results show that the brain employs a trace-rule-like mechanism to associate features across identity-preserving transformations. Wallis and Bühlhoff (2001) showed that subjects will confuse faces which switch identity while undergoing a rotation. Similarly, subjects confuse objects that are consistently swapped when viewed at a particular retinal location (Cox et al., 2005) and AIT neurons change preferred features at the swapped location (Li and DiCarlo, 2008). Li and DiCarlo also performed the analogous experiment for scale invariance (Li and DiCarlo, 2010) showing that the scale invariance of AIT neurons can be broken by consistently changing an object identity during scaling. If as these experiments suggest, the brain employs a trace-rule-like mechanism to build up invariant object representations then we are left with the puzzle of explaining initial invariance for novel objects.

5.4 A role for hierarchical architectures

It turns out that hierarchical models of object recognition provide a resolution to the puzzle of initial invariance. Associating full images of objects presented at different positions in order to build up a translation invariant representation for each object would be unable to explain initially invariant recognition of novel objects. However, associating the object's component parts at different positions is an explanation of how translation invariant representations are learned that is not stymied by the puzzle of initial invariance. Or, in the context of the model discussed in part 3 of this report, features common to a wide range of objects may be associated across positions and used to recognize new objects containing the same or similar features.

Associating parts of objects rather than entire objects at different positions provides a mechanism for resolving the puzzle of initial invariance. Our astronaut may not have seen each Martian

³These studies actually argue for imperfect initial invariance because they observe small drops in performance away from the trained location. However, their results are not inconsistent with there being significant initial invariance for novel objects. Performance remains far higher than chance at untrained locations (Dill and Fahle, 1997; Nazir and O'Regan, 1990). Dill and Edelman (2001) also suggests that there could be two learning processes at work in these studies- one fast invariant process (our initial invariance) and another slower position dependent-effect.

object at every position; however, she would have seen many of the object's component parts at each location. Recognition of the entire object despite shifts in position is accomplished by decomposing the object into features which are themselves detected invariantly of position. The initial invariance to novel objects is thereby inherited from learned invariance to the object's parts.

Continuing with our astronaut on Mars example. We have explained the astronaut's initial lack of visual confusion by appealing to the similarity of features of Martian objects to features of familiar terrestrial objects. It seems reasonable to assume this similarity is real; after-all, natural objects on any planet ought to be composed of lines and edges. It is likely that even more complex intermediate visual features would also be preserved across the two planets. This seems to be a plausible assumption to make. However, it immediately begs a stronger version of the same question: If instead of merely going to Mars, our astronaut had fallen through a wormhole and arrived in an alternate dimension where objects looked to her, like random dot-patterns unlike any object on earth, would she have still been able to consistently recognize objects despite their translation on her retina?

Hierarchical models of object recognition represent objects by their similarity to a set of template features. These template features could be learned by an associative mechanism like the trace rule. However, in the example of our stranded astronaut, no similar earth-bound objects to those on which she would be tested in her new environment were available to learn invariant template features. Would she still invariantly recognize novel alien objects? More concretely, we can ask the questions addressed in the next two sections of this report.

5.5 Do templates need to be learned from similar objects?

Let us return to the aforementioned space-explorer's lack of visual confusion on mars. In her previous earth-bound life, our astronaut had encountered many objects. A lifetime of visual experience with terrestrial objects is likely to be more than enough time to have seen a wide variety of visual features at every location in her visual field. However upon arrival on mars, she was met with a set of entirely new objects, few of which she had seen previously. We hypothesized that the puzzle of her ability to recognize these novel objects despite their translation on her retina is resolved by her prior experience with shared features also present in terrestrial objects.

It is not immediately obvious how far this logic goes. How similar would novel objects need to be to the objects from which the set of invariant features were learned in order to support invariant recognition? Could invariant features extracted from randomly generated dot patterns looking like no real object be sufficient to support recognition of natural objects?

As evident from figure 9 (bottom right panel), even invariant features extracted from highly unnatural objects (random dot patterns) are sufficient to support invariant face identification. There are two⁴ 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 presence of invariant features. Performance decrements attributable to the latter variability are due to the extent

⁴Actually 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 gives 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.

which the feature representation accurately portrays identity without being misled by incidental aspects of the images.

Learning invariant features from similar images to those in the test set helps in producing a feature set that better reflects the most diagnostic aspects of the test images and thus affects overall accuracy level (Serre et al., 2007b). However, the translation invariance itself 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. Our stranded astronaut may have difficulty distinguishing novel Martian objects from one another, but if she can succeed at one eccentricity she will be able to do the same at another.

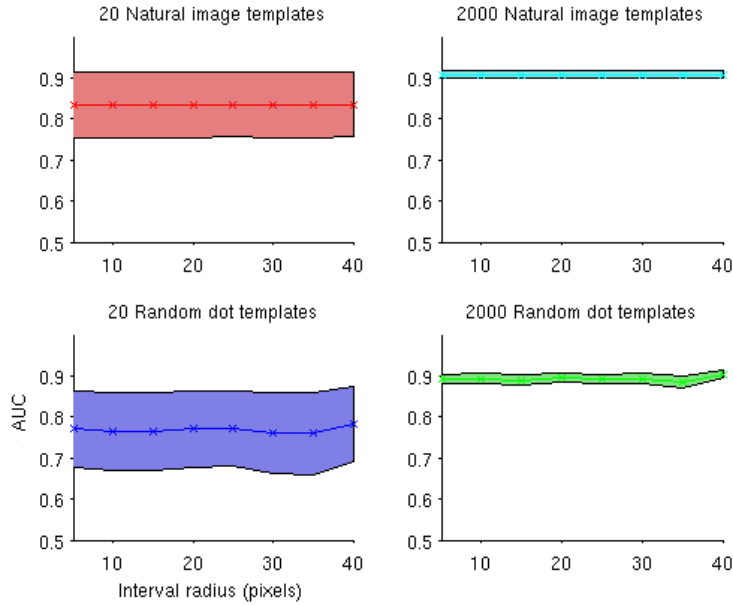


Figure 9: Efficiency of invariance for recognizing targets versus distractors appearing anywhere within an interval of a particular radius (Using C2-layer features only). Red curve: 20 translation invariant features were extracted from natural images. Cyan curve: 2000 invariant features extracted from natural images. Blue and green curves: 20 and 2000 translation invariant features extracted from random dot patterns respectively. We build invariant intermediate features by sampling randomly chosen patches from the template set of images. Error bars display \pm one standard deviation. The test images here were the same as those used in figure 4- right panel.

5.6 How many templates are necessary?

After establishing that invariant 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 9 (left panels) displays the results of two simulations we ran utilizing only a very small number of invariant feature detectors (20) drawn from images of either random dot patterns (pink curve) or natural images (red curve). Surprisingly, these results show that translation invariant recognition is maintained despite very small numbers of features.

When small numbers of features are employed, efficiency is more affected by the suitability of

the features for the particular test images. That is, if you only have a few invariant features it is helpful to have extracted them from similar objects to the test set. Translation invariance however, is unaffected. Figure 10 shows efficiency of invariance (AUC) as a function of the number and type of invariant feature detectors employed.

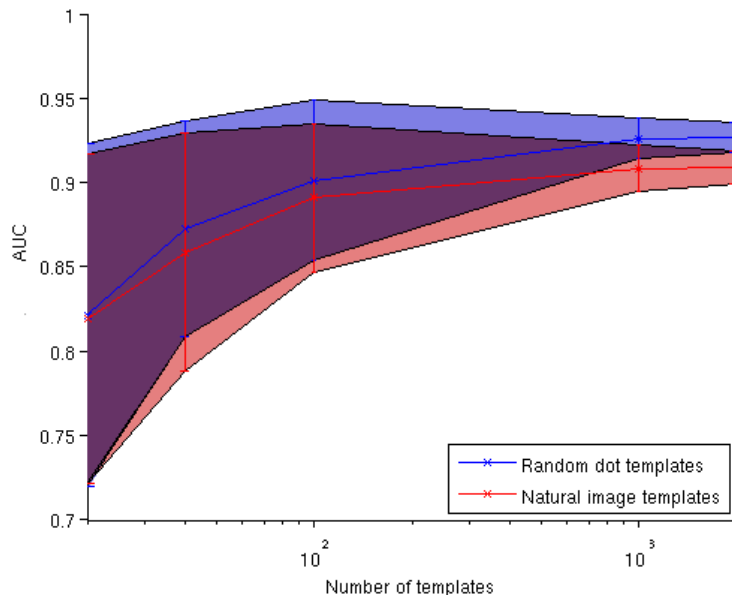


Figure 10: Efficiency of invariance (AUC) for the face identification task as a function of the number of templates used to build the model. Here invariant features were derived from random patches of 20 natural images (red curve) or 20 random dot patterns (blue curve). Restarting the simulation and 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 4B.

The test images used in figures 4B, 6 and 7 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 test 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) Figure 11.

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 chance⁵. Accuracy on this more difficult task increases when we increase the number of templates employed.

⁵Chance would be less than 1% correct on the Caltech101 set of images.

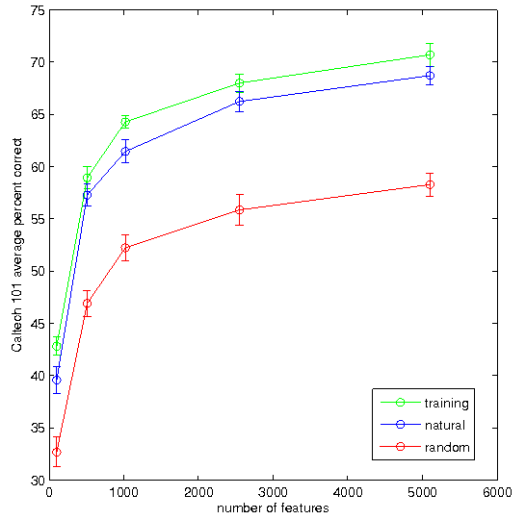


Figure 11: Average percent correct on the Caltech 101 dataset using a regularized least squares (RLS) classifier. with 30 images per class. The templates were extracted from random patches drawn from either the RLS training images (green curve), an unrelated set of natural images (blue curve) or random 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.

5.7 Bootstrapping invariant object recognition

Our purpose here was not to claim that our models are uniquely able to solve the puzzle of initial invariance. Rather, our goal was to demonstrate that a general computational principle- associative mechanisms operating on units that detect template features of intermediate complexity- is sufficient to solve the problem. This result is compatible with a broad class of hierarchical models of visual object recognition (Fukushima, 1980; LeCun et al., 1989; Mel, 1997; Perrett and Oram, 1993; Riesenhuber and Poggio, 1999). It is also agnostic to the particular mechanism by which template features become associated with one another.

Toward that end, we had two unexpected results.

- 1). The objects from which the template features are learned need not be similar to the objects on which the system will be tested. Translation invariance for novel objects follows from the presence of translation invariant features. It is not necessary that the new objects be well-represented by the invariant features.
- 2). The set of invariant features does not even need to be very large in order for novel objects to be recognized despite translation. In fact, a very small ~ 10 number of features was shown to be sufficient.

Taken together, these two results have important implications for the efficiency of learning invariant object representations. Recall that one of the seemingly most difficult points for the memory-based approach is the apparent need to have previously seen large numbers of objects at every position in the visual field. These results show that in fact, only a very small number of objects need to have been seen at all positions. Moreover, these objects could be from any

class at all and their features would still be able to support invariant object recognition for novel objects.

These results imply a model by which an infant could rapidly learn invariant representations for a wide variety of objects. 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.

A newborn infant could acquire a relatively small set of invariant features through association of convenient transforming features in its environment. We may call these features the primal templates. Alternatively, these features may be learned from spontaneous retinal waves. When learning to recognize new objects, the infant could rely on the initial invariance inherited from the primal templates. On its own this would not be enough to guarantee high accuracy at recognizing new objects; indeed accuracy with new objects would be determined by how well the primal templates happens to represent the object.

However, utilizing the primal templates, the infant's visual system has additional information by which additional features can be incorporated into invariant object representations. The weakly activated (but invariant) response of the primal templates could serve as a cue that the object has undergone an identity-preserving transformation and so whatever other patterns were also evoked by the object across the transformation should also become associated. With such a mechanism, the brain could bootstrap from an initially simple set of primal templates into a much more complex system capable of representing the full richness of the visual world.

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.
- 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 Cox, D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

- 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.
- 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.
- 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.
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.
- 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. 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.
- Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.
- 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.
- Mel, B. W. (1997). SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition. *Neural Computation*, 9(4):777–804.

- 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.
- Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333.
- Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harb Symp Quant Biol*.
- Poggio, T. and Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, 431(7010):768–774.
- Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–266.
- Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.
- 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.
- 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 Bühlhoff, H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771.
- Wallis, G. and Bühlhoff, 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.
- 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.