Visual Cortex and Deep Networks

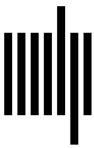
Learning Invariant Representations

By: Tomaso A. Poggio, Fabio Anselmi

Citation:

Visual Cortex and Deep Networks: Learning Invariant Representations
By: Tomaso A. Poggio, Fabio Anselmi
DOI: 10.7551/mitpress/10177.001.0001
ISBN (electronic): 9780262336710
Publisher: The MIT Press
Published: 2016

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

3 Retinotopic Areas: V1, V2, V4

3.1 V1

3.1.1 A New Model from i-Theory for Eccentricity Dependence of Receptive Fields

The simplest and most common image transformations are affine transformations and in particular the subgroup consisting of shifts in position (x) and uniform changes in scale (s). The theory suggests that the first step of the invariance computation likely consists of learning the set of transformations. This is implicitly done via actual visual experience (or by evolution encoded in the genes or by a combination):

1. Store template t^k (which is an image patch and could be chosen at random).
2. Store all the observed transformed templates (bound together by continuity in time).
3. Repeat steps 1 and 2 for a set of K templates.

Although the templates can be arbitrary image patches, we assume—because of the optimal invariance theorem (see chapter 1) and because of the experimental evidence from V1 simple cells—that V1 templates are Gabor-like functions (more precisely, windowed Fourier transforms [72], p. 92).

Under scaling, a pattern exactly at the center of the fovea will change size without any translation of its center, and its boundaries will shift in x, y. For a pattern centered at some nonzero eccentricity, scaling will translate its center in the s, x plane (see figure 3.1). In the s, x plane the slope of the trajectory of a pattern under scaling is a straight line through the origin with a slope$s$ that depends on the size of the pattern and the associated position.

Consider a Gabor function at $s = s_0, x = 0$. s_0 is the minimum possible receptive field (RF) size, given optical constraints. Transforming it by shifts in x within $(-x_0, x_0)$ generates a set Λ_0 of templates. Suppose we want to ensure that what is recognizable at the highest resolution (s_0) remains recognizable at all scales up to $s_{\max}$.

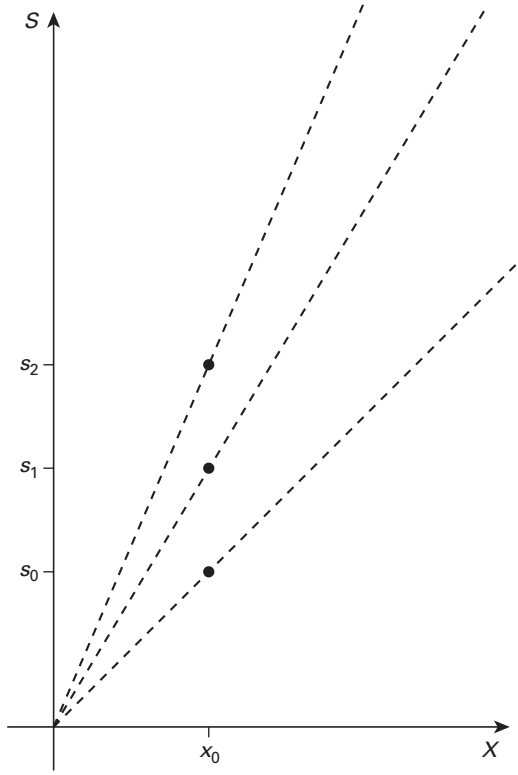


Figure 3.1

Left, the s, x plane, indicating templates of radius s_1, s_2, and s_3 in the spatial dimensions x, y, all centered at position X. While s and x are both measured in degrees of visual angle, in this plot the two axes are not shown to the same scale. Here, as in the rest of the book, we show only one spatial coordinate (x); everything we say can be directly extended to the x, y plane. We assume that the smallest template is the smallest simple cell in the fovea with a radius of around $40''$ [73]. For any fixed eccentricity (*right*) the size of the pattern determines the slope of its s, x trajectory under scaling. *Source:* [14].

The associated scale transformations of the set Λ_0 yield the inverted truncated pyramid shown in figure 3.2. Pooling over that set of transformed templates according to the Hubel-Wiesel algorithm (see chapter 1) will give uniform invariance to all scale transformations of a pattern over the range $(s_0, s_{\max})$; invariance to shifts will be at least within $(-x_0, x_0)$, depending on scale. The process of observing and storing a set of transformations of templates in order to be able to compute invariant representations may take place at the level of evolution or at the level of development of an individual system or as a combination of both.

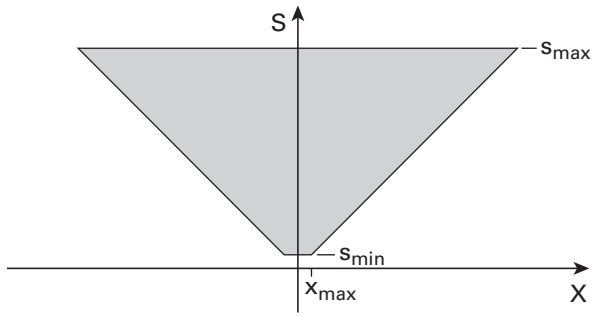


Figure 3.2

Synthesis of template set. Consider a template, such as a Gabor receptive field (RF), at $s_{\min}$ and $x = 0$. Store its transformations under (bounded) shift, filling the interval between $s_{\min}$, $x = 0$ and $s_{\min}$, $x_{\max}$. Store its transformations over (bounded) scale, filling the space in the inverted truncated pyramid. This is the space of bounded joint transformations in scale and space. For clarity, axes s, x are not in the same units (s unit is 10 times the x unit, so the real slope is $1/10$ of the slope shown). *Source:* [14].

The following definition holds. The inverted truncated pyramid of figure 3.2 is the locus Λ of the points such that their scaling between $s_{\min}$ and $s_{\max}$ gives points in S; further, all points P in $(-x_0, x_0)$ are in S, for example, $P \in \Lambda$ if

$$g_s P \in \Lambda, \quad s \in (s_{\min}, s_{\max}), \quad \Lambda_0 \in \Lambda, \tag{3.1}$$

where Λ_0 consists of all points at $s_{\min}$ between $-x_0$ and x_0. Other alternatives are possible: one is to set a constant difference between the minimum and the maximum scale at each eccentricity: $s_{\max} - s_{\min} = \text{const}$. Experimental data suggest this is a more likely possibility (large eccentricities are also represented in the visual system).

The new model consists of an inverted pyramid region. It follows naturally from i-theory if scale invariance has a higher priority than shift invariance. Recall that according to the optimal invariance theorem, the optimal template is a Gabor function. Under scale and translation within the inverted pyramid region, the Gabor template originates a set of Gabor wavelets (a tight frame). In the region of figure 3.2 scaling each wavelet generates other elements of the frame.

Note that the shape of the lower boundary of the inverted pyramid, although similar to standard empirical functions that have been used to describe the cortical magnification factor (M) scaling, is different for small eccentricities. An example of an empirical function [74, 75] is $M^{-1} = M_0(1 + ax)$, where M is the cortical magnification factor, M_0 and a are constants, and x is eccentricity.

A normal, nonfiltered image may activate all or part of the Gabor filters within the inverted truncated pyramid of figure 3.2. The pattern of activities is related to a multiresolution decomposition of an image. We call this transform of the image an inverted pyramid scale-space fragment (IP fragment) (the term *fragment* in this regard is borrowed from Ullman [76]) and consider it as supported on a domain in the space x, y, s that is contained in the inverted truncated pyramid of the figure. The fragment corresponding to a bandpass filtered image should be a more or less narrow horizontal slice in the s, x plane. In the following we assume that the template is a Gabor filter (of one orientation; other templates may have different orientations). We assume that the Gabor filter and its transforms under translation and scaling are roughly bandpass, and the sampling interval at one scale over x is s, implying half overlap of the filters in x. This is illustrated in figure 3.3. These assumptions imply that for each array of filters of size s, the first unit on the right of the central one is at $x = s$, if x and s are measured in the same units. For the scale axis we follow the sampling pattern estimated by Marr [73] with five frequency channels with $2s = 1'20'', 3.1', 6.2', 11.7', 21'$. Filter channels as described are supported by sampling by photoreceptors that starts in the center of the fovea at the Shannon rate, dictated by the diffraction-limited optics with a cutoff around 60 cycles/degree, and then decreases as a function of eccentricity.

3.1.2 Fovea and Foveola

In this truncated pyramid model of the simple cells in V1, the slope of the magnification factor M as a function of eccentricity depends on the size of the foveola, which we define here as the region at the minimum scale $s_{\min}$. The larger the foveola, the smaller the slope. We submit that this model nontrivially fits data about the size of the fovea, the slope of M, and the size of receptive fields in V1. In particular, the size of the foveola, the size of the largest RFs in the anterior in ferotemporal cortex (AIT), and the slope of acuity as a function of eccentricity depend on each other. Fixing one determines the other (after setting the range of spatial frequency channels, i.e., the range of RF sizes at $x = 0$ in V1). For a rough calculation we assume here that $s_{\min} \approx 40''$ (from an estimate of $1'20''$ for the diameter of the smallest simple cells [73, 77]). Data from Hubel and Wiesel [20, 78, figure A.6] and Gattass [79, 80], shown in figure 3.4, yield an estimate of the slope a for M in V1 (the slope of the lines $s_{\min}(x) = ax$). Hubel and Wiesel cite the slope of the average RF diameter in V1 as $a = 0.05$; Gattass quotes a slope of $a \sim 0.16$ (in both cases the combination of simple and complex cells may yield a biased estimate relative to the true slope of simple cells alone). Based on these actual data, our model of an inverted truncated pyramid

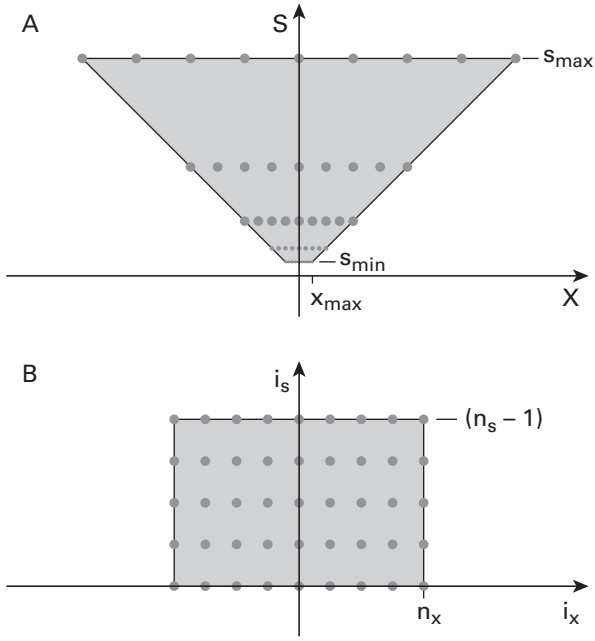


Figure 3.3

Under the assumption of Gabor filters and associated sampling for each scale at spatial intervals $\Delta x = s$, this figure depicts a subset of the resulting array of template units. The sampling over x follows the sampling theorem. There are no samples between the s axis and the line with slope 1 (when x and s are plotted in the same units). The center of the circles gives the s, x coordinates; the circles are icons symbolizing the receptive fields. The figure shows a foveola of $\approx 26'$ and a total of ≈ 40 units (20 on each side of the center of the fovea). It also shows sampling intervals at the coarsest scale in the fovea (assumed to be around $2s = 21'$ [73], which would span $\approx \pm 6$. The size of a letter on the ophthalmologist's screen for 20/20 vision is around $5'$. Since the inverted truncated pyramid (A) has the same number of sample points at every scale, it maps perfectly onto a square array (B) when x is replaced by $i_x = x/s$, the number of samples from the center. i_s is the scale band index (s, x units are scaled as in figure 3.2 for clarity). *Source:* [14].

predicts (using an estimate of $a = 0.1$) that the radius of the foveola (the bottom of the truncated pyramid) is $R = 1'20''/0.1 \approx 13'$ with a full extent of $2R \approx 26'$ corresponding to about 40 cells separated by $40''$ each. The size of the fovea (the top of the truncated pyramid) is predicted to then have $2R \approx 6'$ with 40 cells spaced $\approx 10'$ apart (see figure 3.3). Each of the scale layers has the same number of units, which is determined by the number of units in the foveola, that is, the number of units at the finest resolution. This remapping (see figure 3.3B) shows that $S1$ corresponds to a lattice of dimensions x, y, s, θ, where the dimension sizes are different (but roughly the same for x, y); the topology is that of a cylinder, with θ periodic.

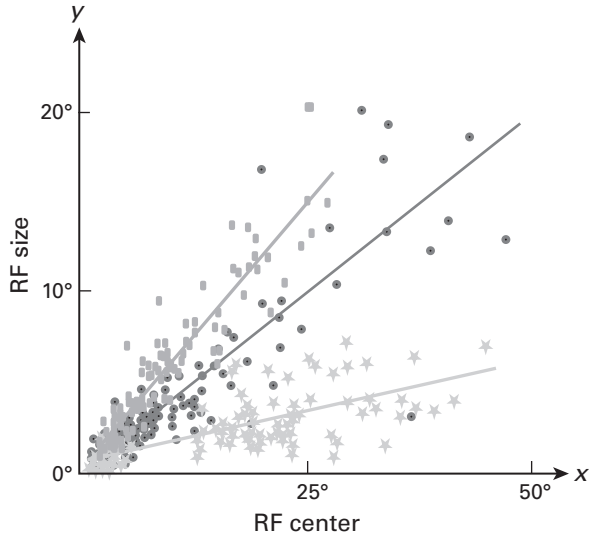


Figure 3.4

Data from Hubel and Wiesel [78] for monkey V1 give a slope for average RF diameter, relative to eccentricity, of $a = 0.05$. Data from other areas are similar but have higher slope (adapted from [48] with original monkey data from [79, 80]). *Source:* [14].

The lattice contains all the sampled translations and scale transformations of the templates required for invariance within the inverted pyramid. The number predicted by the theory is large but quite reasonable: $40 \times 40 = 1,600$ per scale and per Gabor orientation for a total of around 50K transformed templates.

To make these predictions, we used data from the macaque together with data from human psychophysics. Thus our estimates depend on the actual range of receptive field sizes and could easily be wrong by factors of 2. Our main goal is to provide a ballpark estimate of relevant quantities.

3.1.3 Scale and Position Invariance in V1

Invariance is not provided directly by the array but by the pooling over it. We limit ourselves in this section to the invariant recognition of isolated objects. The range of invariance in x is limited for each s by the slope of the lower bound of the inverted pyramid. The prediction is that the range of invariance Δx is proportional to the scale s, that is,

$$\Delta x \approx n_x s, \tag{3.2}$$

where n_x is the radius of the inverted pyramid and is the same for all scales s. Thus small details (high frequencies) have a limited invariance range in x,

whereas coarser details have larger invariance. n_x is obtained from the slope a of the cortical magnification factor as

$$n_x = \frac{1}{a}. \tag{3.3}$$

The following rationale for a so-called normative theory seems natural (and could easily be wrong):

1. Evolution tries to optimize recognition from few labeled examples,
2. As a consequence it has to optimize invariance,
3. As a first step it has to optimize invariance to scale and translation (under constraint of noninfinite resources),
4. Therefore it develops in V1 multiple sizes of receptive fields at each position, with RF size increasing linearly with eccentricity, properties that are reflected in the architecture of the retina.

3.1.4 Tuning of Cells in V1

Chapter 2 described the learning of templates through unsupervised experience of their transformations. The main example discussed consists of simple cells in V1. We can now add details to the example using the inverted pyramid architecture. Our basic hypothesis is that the position and size (a Gaussian distribution) of each immature simple cell is set by development and corresponds to a node in the lattice of figure 3.3 in x, y, s. Furthermore, Hebbian synapses on the simple cells will drive the tuning of the cell to be the eigenvector with the largest eigenvalue of the covariance of the input images. It can be shown ([11] and appendix section A.6) that because of the Gaussian envelope at each site in the lattice and because of the statistics of natural images, the tuning of each simple cell will converge to a Gabor-like function with properties that match very closely the experimental data on the monkey, the cat, and the mouse (see figure 1.4). Since the arguments in chapter 2 apply directly to this case, pooling over simple cells is equivalent to pooling over transformations of a template (a Gabor function with a specific orientation).

3.2 V2 and V4

3.2.1 Multistage Pooling

As discussed, pooling in one step over the whole s, x domain of (figure 3.3) suffers from interference from clutter-induced fragments in any location in the inverted pyramid. A better strategy is to pool area by area (in the ventral

stream), that is, layer by layer (in the corresponding hierarchical architecture). The *C* cells' activities in each layer of the hierarchy are sent to the memory or classification module at the top. After each pooling (*C* unit) stage, and possibly also after a dot product (*S* unit) stage, there is a downsampling of the array of units (in x, y and possibly in s) that follows from the low-pass-like effect of the operation. Note that according to i-theory, the templates in V2 and V4 should be patches of neural images at that level, possibly determined by PCA-like learning.

Here we analyze the properties of a specific hypothetical strategy of downsampling by 2 in x, y at each stage. We focus on shift invariance because the number of scales is so small (≤ 5) that pooling in scale could even be avoided all together. The choice of downsampling by 2 in each spatial dimension is simple and seems consistent with biological data. It is easy to modify the results by using the same logic with a different criterion for downsampling. Pooling each unit over itself and its neighbors (thus a patch of radius 1) allows downsampling of the array in x by a factor of 2. We assume that each combined *S*-*C* stage brings about a downsampling by 2 in each dimension of the x, y array. We call this process decimation; see figure 3.5.

Starting with V1, four stages of decimation reduce the number of units at each scale from ≈ 40 to ≈ 2, spanning $\approx 26'$ at the finest scale and $\approx 6'$ at the coarsest. Pooling over scale in a similar way may also decimate the array down to about one scale from V1 to IT. Neglecting orientations, in x, y, s the $30 \times 30 \times 6$ array of units may be reduced to just a few units in x, y and one in scale. This picture is consistent with the invariance found in IT cells [38]. According to i-theory, different types of such units are needed, one for each of several templates at the top level. Other types of pooling and downsampling are conceivable, such as pooling in space over each spatial frequency channel separately, possibly in different areas. Understanding the actual strategy underlying hierarchical pooling in the ventral stream is an open question. The simple strategy we describe is just a conjecture. Notice that downsampling in space by 2 at each stage, together with our previous estimates of the dimensions of the inverted pyramid in V1, predicts that about four to five layers in the hierarchy are required to obtain almost full invariance to shift and scale before stage 2. If layers are identified with visual areas, there should be at least four areas from V1 to AIT: it is tempting to identify them with V1, V2, V4, PIT.

3.2.2 Predictions of Crowding Properties in the Foveola and Outside It (Bouma's Law)

The pooling range is uniform across eccentricities in figure 3.6. The spatial pooling range depends on the area: for V1 it is the sampling interval between

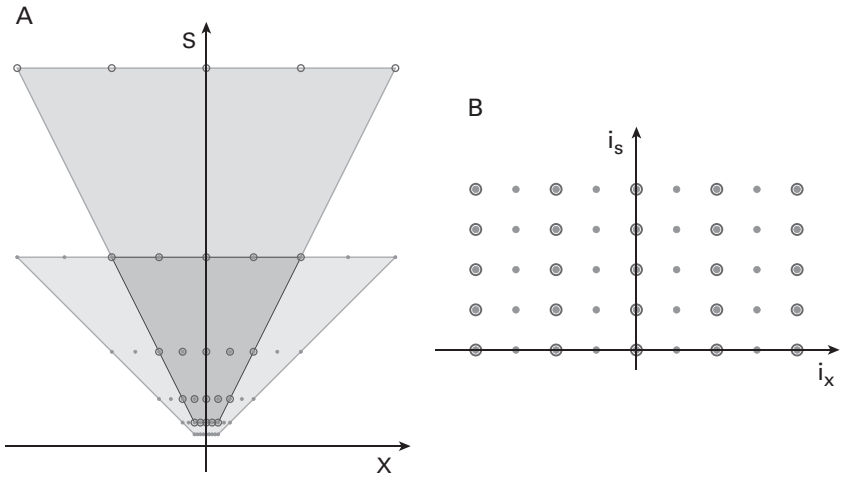


Figure 3.5

Pooling over 3×3 patches in x, y of the lattice of simple cells in V1 and subsampling decimates the lattice. If the lattice in x, y is 40 units, four steps (V1, V2, V4, PIT) of x pooling are sufficient to create cells that are 16 times larger than the largest in the fovea in V1 (probably around $21'$ at the coarsest scale), yielding cells with an RF diameter of up to ≈ 5. Each area in the fovea would see a doubling of size with corresponding doubling of the slopes at the border (before remapping to a cube lattice). The index of the units at position x and scale s is given by $i_x^s = 2^s i_x^1$. Simultaneous pooling over regions in S is possible. *Source:* [14].

the dots, for V2 it is the sampling interval between the shaded dots: they should be roughly equivalent to the radius of the receptive field of the complex cells in V1 and V2, respectively. V4 is not shown, but it is clear what is expected.

Figure 3.6 shows the regions of pooling corresponding to our assumptions. The inverted pyramid is split into sections, each one corresponding to pooling by a complex cell module. For now we consider only pooling in space and not scale. Figure 3.6 describes the situation for V1, but the same diagram can also be used for V2, V4, and PIT by taking into account the downsampling, the increase in size of the smallest cells, and the doubling in slope of the lower boundaries of the inverted pyramid. It is clear from the figure why the following criterion for pooling to remain interference-free from a flanking object, that is, unaffected by clutter, seems reasonable: the target and the flanking distractor must be separated by at least the pooling range and thus by a complex cell receptive field.

We consider two cases: (1) the target is in the central section at some layer (say, V2), and (2) the target is outside the foveola, that is, at an eccentricity greater than $10'$. Our predictions are as follows:

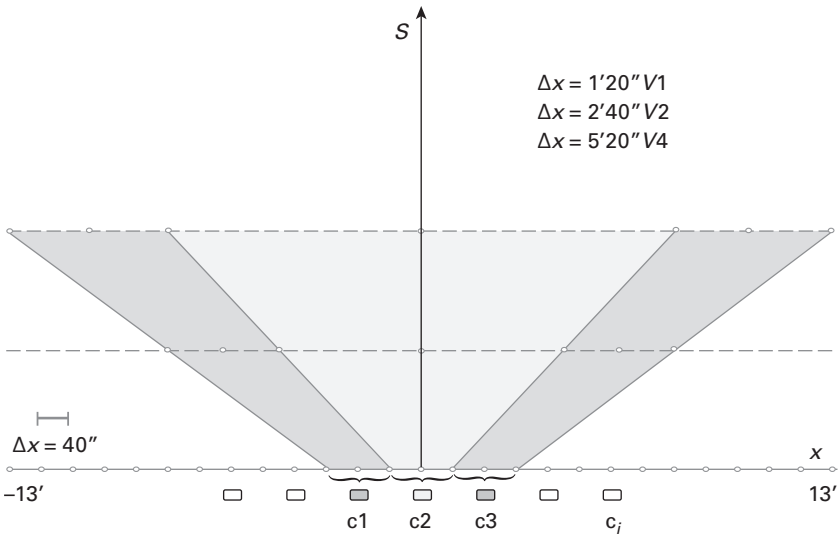


Figure 3.6

Part of the inverted pyramid of simple units in V1. The shaded regions are pooling regions, each one corresponding to a complex cell (there are more such regions to cover the extent of the foveola (diameter $\approx 21'$). The ones at the left and right boundaries will be at the edges of the inverted pyramid. Pooling shown here is spatial only (at each scale); it takes place around each simple unit and its two neighbors in x, y (a 3×3 patch). *Source:* [14].

- Consider a small target, such as a $5'$ width letter, placed in the center of the fovea, activating the smallest simple cells at the bottom of the inverted pyramid. The smallest critical distance to avoid interference should be the size of a complex cell at the smallest scale, that is, $\Delta x \approx 1'20''$ in V1 and $\Delta x \approx 2'40''$ in V2. Of course, in general, more than a scale will be activated. In particular, if the letter is made larger, the activation of the simple cells shifts to larger scales (s in figure 3.6), and since $\Delta x \approx s$, the critical spacing is proportional to the size of the target. It is remarkable that both these predictions match quite well figure 10 in Levi and Carney [81], although their data are for orientation discrimination.
- Outside the foveola, the size of the relevant complex cells determining pooling depends on spatial frequency content of target and distractors and on whether the pooling is only over space or also over scales (likely). Since the smallest RF size of the complex cells increases linearly with eccentricity, we expect that minimal critical spacing to avoid crowding outside to be roughly $\Delta x \approx bx$, with b depending on the cortical area representing the bottleneck. Thus the theory "predicts" Bouma's law [82] of crowding (often Bouma's

law is given as $\Delta x \approx bx + w$; see [83–85]). The experimental value found by Bouma for crowding of $b \approx 0.5$ points to a retinotopic area in the ventral stream, which may be V4, since the smallest slope found by Gattass for dependence on eccentricity of RF size is for V1 ≈ 0.12, for V2 ≈ 0.25, and for V4 ≈ 0.5 (on the other hand, studies of metameric stimuli by Freeman and Simoncelli [48] were interpreted by them to imply V2 in crowding and peripheral vision deficiencies).

3.2.3 Scale and Shift Invariance Dictates the Architecture of the Retina and the Retinotopic Cortex

It is interesting that from the perspective of i-theory, the linear increase of RF size with eccentricity, found in all primates, follows from the computational need of computing a scale- and position-invariant representation for novel images/objects. The theory also predicts the existence of a foveola and links its size to the slope of linear increase of receptive field size in V1. From this point of view, the eccentricity dependence of cone density in the retina as well as of the cortical magnification factor follow from the computational requirement of invariant recognition. The usual argument of limited resources (say, number of fibers in the optical nerve) does not determine the shape of the inverted pyramid but only the size of the fovea at the bottom (and thereby of the total number of cells). The inverted pyramid shape is independent of any bound on computational resources.

As mentioned, the estimate for the size of the foveola is quite small, with a diameter of $26'$ corresponding to about 40 simple cells (of each orientation) and about 50 cones in the retina. This is almost certainly a trade-off between limited translation invariance (the inverted pyramid region) and the ability to correct for it by moving gaze, while scale invariance is fully available in the center of the fovea. Notice that a fovea with a 10 times larger diameter would lead to an optical nerve of the same size as the eye, making it impossible to move it.

This state of affairs means that there is a quite limited field of vision in a single glimpse. Most of an image of 5×5 degrees is seen at the coarsest resolution only, if fixation is in its center. At the highest resolution only a very small fraction of the image (up to $30'$) can be recognized, and an even smaller part of it can be recognized in a position-invariant way (these numbers are rough estimates). An IP fragment can be defined as the information captured from a single fixation of an image. Such a fragment is supported on a domain in the space x, s, contained in the inverted truncated pyramid of figure 3.2. For normal-sized images and scenes with fixations well inside, the resulting

IP fragment will occupy most of the spatial and scale extent of the inverted pyramid.

Consider now the fragment corresponding to an object to be stored in memory (during learning) and recognized at runtime. Consider the most favorable situation for the learning stage: the object is close to the observer so that both coarse and fine scales are present in its fragment. At runtime the object can be recognized whenever it is closer or farther away (because of pooling). The important point is that a look at this and the other possible situations (at the learning stage) suggests that the matching should always weight more the finest available frequencies (bottom of the pyramid). This is the finding of Schyns and Gosselin [86]. As implied by their work, top-down effects may modulate these weights somewhat (this could be done during pooling), depending on the task. Assume that such a fragment is stored in memory for each fixation of a novel image. Then, because of large and position-independent scale invariance, there is the following trade-off between learning and runtime recognition:

- If a new object is learned from a single fixation, recognition may require multiple fixations at runtime to match the memory item (given its limited position invariance and unless fixation is set to be within the object).
- If a new object is learned from multiple fixations, with different fragments stored in memory each time, runtime recognition will need a lower number of fixations (in expectation).

The fragments of an image stored in memory via multiple fixations could be organized into an egocentric map. Though the map may not be directly used for recognition, it is probably needed to plan fixations during learning and especially during the recognition and verification stages (and thus indirectly used for recognition in the spirit of minimal images and related work by Ullman and coworkers,personal communication). Such a map may be related to Marr's $2\frac{1}{2} = D$ sketch [87], for which no neural correlate has been found as yet.

Thus, simultaneous invariance to translation and scale leads to an inverted truncated pyramid as a model for the scale-space distribution of the receptive fields in V1. This is a new alternative, as far as we know, to the usual smooth empirical fit of the cortical magnification factor data. In particular, the linear dependence on eccentricity of receptive field sizes in V1 (and cones sampling in the retina) follows from the computational requirement of scale invariance. This picture contains other interesting properties: the range of shift invariance depends on scale; the size of the flat, high acuity foveal region, which we identify with the foveola, can be inferred from the slope of the eccentricity-dependent acuity. Existing neural data from macaque V1 suggest a foveola with a diameter of around $20'$ of arc. In a sense, scale invariance turns out to

be more natural than shift invariance in this model (there is scale invariance over the full range at the fixation point). This is to be expected in organisms in which eye fixations can easily take care of shift transformations, whereas more extensive motions of the whole body would be required to change the scale. i-theory predicts scale invariance at this level: this is exactly what Anstis [87] found: a letter just recognizable at some eccentricity remains equally recognizable under scaling.

Physiology data suggest that AIT neurons are somewhat less tolerant to position changes of small stimuli [88, 89]. A comparison across studies suggests that position tolerance is roughly proportional to stimulus size [89]. If we assume that some IT neurons effectively pool over all positions and scales, i-theory in fact expects that their invariant receptive field (over which consistent ranking of stimuli is maintained) should be smaller for higher spatial frequency patterns than for low-frequency ones. There is evidence of attentional suppression of nonattended visual areas in V4. From the perspective of i-theory, it seems natural that top-down signals may be able to control the extent of pooling, or the pooling stage, used in computing a signature from a region of the visual field, in order to minimize clutter interference.

In summary, i-theory provides explanations and computational justifications for several known properties of the retinotopic cortex. It also makes a few predictions that are still waiting for experimental tests:

- There is an inverted pyramid of simple cells size and positions with parameters specified in the text, including linear slope of the lower boundaries. The predicted pyramid is consistent with available data. More precise measurements in the region of the foveola could decide between the usual empirical fits and our predictions.
- Anstis [87] did not take measurements close to the minimum letter size, which is about $5'$ for $20/20$ vision. i-theory predicts that if there is a range of receptive fields in V1 between $s_{\min}$ and $s_{\max}$ in the fovea, then there is a finite range of scaling between $s_{\min}$ and $s_{\max}$ under which recognition is maintained [14]. It is clear that looking at the image from an increasing distance will at some point make it unrecognizable; it is somewhat less clear that getting too close will also make it unrecognizable (this phenomenon was found in Ullman's minimal images [Ullman, personal communication]).
- Consider the experimental use of images such as novel letters (never seen before) of appropriate sizes that are bandpass filtered (with the Gabor-like filters assumed for V1). The prediction—because of the pooling over the whole inverted pyramid (done between V1, V2, and V4) — is that for a new presentation there will be psychophysical scale invariance for all frequencies

between $s_{\min}$ and $s_{\max}$. Shift invariance increases linearly with spatial wavelength and is at any spatial frequency at least between $x_{\min}$ and $x_{\max}$ (the bottom edge of the truncated pyramid).

- i-theory predicts a flat region of constant maximum resolution, which we called the foveola. Its size determines the slope of the lower border of the pyramid. Since the slope can be estimated relatively easily from existing data, our prediction for the linear size of the foveola is about $40'$ of arc, corresponding to about 30 simple cells of the smallest size (assumed to be $\approx 1'20''$ of arc). Note that our definition of the fovea is in terms of the set of all scaled versions of the foveola between $s_{\min}$ and $s_{\max}$ spanning about 6 degrees of visual angle.
- i-theory explains crowding effects in terms of clutter interference in the pooling stage (see also [90]). It predicts Bouma's law and its linear dependence on eccentricity [82]. Since Bouma's constant has a value of about 0.4 (see [48]), our theory requires that a signature that is interference-free from clutter at the level of V2 is critical for recognition. This is consistent with the i-theory independent requirement [8, 9] that signals associated with image patches of increasing size from different visual areas must be able to access memory and classification stages separately. The requirement follows from the need to recognize parts and wholes in an image and to avoid clutter for small objects. The V2 signal could directly or indirectly (via IT, or V4 and IT) reach memory and classification.
- The angular size of the fovea remains the same at all stages of a hierarchical architecture (V1, V2, V4,...), but the number of units per unit of visual angle decreases and the slope increases because the associated x increases (see figure 3.5). The theory predicts crowding in the foveola (very close to fixation) but with very small Δx (see eq. (3.2)) that depends on the size of the (small) objects rather than eccentricity. For objects smaller than $20'$ in diameter, the prediction (to be tested) is $\Delta x \approx 3' - 4'$, assuming that the main effect of clutter is in V2 and at the smallest of the five channels of simple cells.

3.2.4 Tuning of Simple Cells in V2 and V4

It is difficult to make clear predictions about V2 and V4 without additional specific data because several options are allowed by the theory. A simple scenario is as follows. There is an x, y, s lattice for V2 (and another for V3) simple cells, as shown in figure 3.5. The tuning of each simple cell (a point in the x, y, s lattice) is determined by the top principal component computed on the neural activity of the complex cells in V1, seen through a Gaussian window that

includes $\approx 3 \times 3$ complex cells in x, y and a set of orientations for each position and scale. It is likely—and supported by preliminary experiments [11]—that some of the principal components computed in this way over a large number of natural images can lead to cell tunings similar to measurements in V4.

Background and Bibliography

This part of the book describes work in our group. Its main source is the technical report [14].

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.

Library of Congress Cataloging-in-Publication Data

Names: Poggio, Tomaso, author. | Anselmi, Fabio, author.
Title: Visual cortex and deep networks : learning invariant representations / Tomaso A. Poggio and Fabio Anselmi.
Description: Cambridge, MA : MIT Press, [2016] | Series: Computational neuroscience | Includes bibliographical references and index.
Identifiers: LCCN 2016005774 | ISBN 9780262034722 (hardcover : alk. paper)
Subjects: LCSH: Visual cortex. | Vision. | Neural networks (Neurobiology) | Perceptual learning. | Computational neuroscience.
Classification: LCC QP383.15 .P64 2016 | DDC 612.8—dc23 LC record available at http://lccn.loc.gov/2016005774

10 9 8 7 6 5 4 3 2 1