Towards a theory of computation in the visual cortex

David A. Mély & Thomas Serre
Cognitive, Linguistic & Psychological Sciences Department
Brown Institute for Brain Science
Brown University, Providence, RI 02912

One of the major goals in visual neuroscience is to understand how the cortex processes visual information [57]. A substantial effort has thus gone into characterizing input-output relationships across areas of the visual cortex [16], which has yielded an array of computational models. These models have, however, typically focused on one or very few visual areas, modules (form, motion, depth, color) or functions (*e.g.*, object recognition, boundary detection, action recognition, *etc.*), see [73] for a recent review. An integrated framework that would explain the computational mechanisms underlying *vision* beyond any specific visual area, module or function, while being at least consistent with the known anatomy and physiology of the visual cortex is still lacking.

The goal of this review is to draft an initial integrated theory of visual processing in the cortex. We highlight the computational mechanisms that are shared across many successful models and derive a taxonomy of canonical computations. Such an enterprise is reductionist in nature as we break down to a basic set of computations the myriad of input-output functions found in the visual cortex. Identifying canonical computations that are repeated and combined across visual functions will pave the way for the identification of their cortical substrate [5].

Canonical microcircuits are a theorist's dream because their very existence provides evidence that different visual cortices indeed tackle one common set of computational problems with a shared toolbox of computations, and constitute the building blocks of the different information processing pathways throughout the visual cortex. While the identification of canonical circuits has proved elusive [17], recent technological advances hold new promises: from complete anatomical reconstructions of circuits in increasingly large volumes of brain tissue [72] to the recording of increasingly large populations of neurons [99] combined with an ability to selectively modulate their activity [20].

We start this review with an overview of cortical filter models, focusing on the two-dimensional Gabor function, a notorious example successfully used to characterize the receptive field of orientation-selective cells as found in the primary visual cortex. We use this model as a case in point for understanding

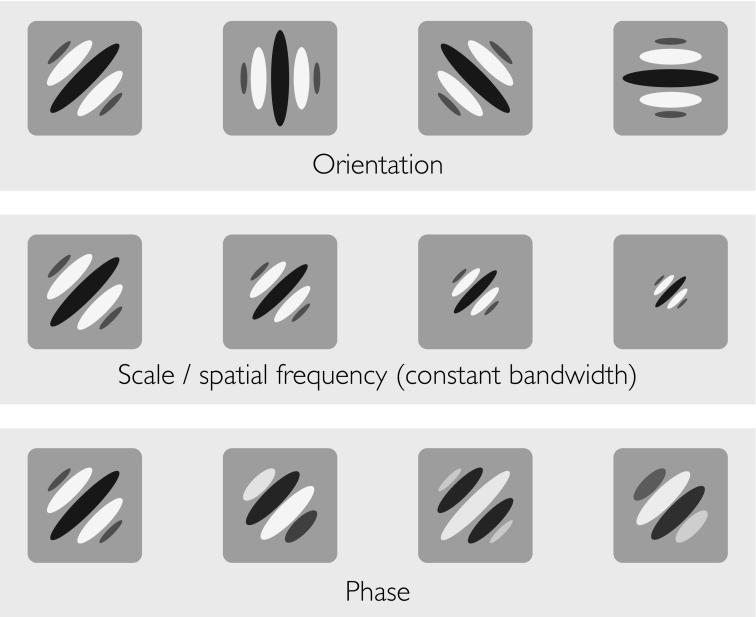

Figure 1: **Cortical filter models of the primary visual cortex**. A good linear filter model that can be well fitted to simple cells is the Gabor filter [55, 11, 12, 43]. (Note that other parametrizations are also possible, see text.) Computational models of the primary visual cortex typically include a battery of such filters spanning a range of orientations, spatial frequencies and phases.

how computational mechanisms for the processing of two-dimensional shape extend to other domains including color, binocular disparity, and motion. We further demonstrate how these basic computations can be cascaded within a hierarchical architecture to account for information processing in extra-striate areas, examining the theoretical underpinnings of their effectiveness. Finally, we describe other possible canonical circuits or principles of design that hold promise for future research.

1 Cortical filter models of form processing

1.1 The linear-nonlinear (LN) model

Cortical filter models, also referred to as "cortex transform" models [105], have been widely used to describe the input-output transfer function of neurons across cortical areas and visual functions (see [48] for a broad overview). Under this broad family of models, the output (or activity, by reference to the firing of biological neurons) of a model cell (also called "unit") depends on the activity of units that feeds into it (the "afferent" units), which is called its "receptive

field". In turn, a particular unit may project onto a set of output units which are called "projection" units. Any afferent unit may itself have its own input units; in the case of vision, such cascades can be traced all the way back to the retina. Thus, by extension, the receptive field of a unit also designates the unique sub-region of the visual field that if properly stimulated may elicit a response from the unit.

Well before the advent of modern computational modeling, neurophysiologists had developed the tools to map out the input-output function of cortical cells in the primary visual cortex. One prominent experimental method, derived from systems theory, is known as "reverse correlation" (see [82] for a review): a neuron is treated as a black box which transforms a visual input $\mathbf{x}$, *i.e.*, the set of image elements (or pixels) $x_{i,j}$ for i, j in its receptive field, into an output response y. The neuron's input-output relationship is characterized as a linear function of its input given by the following equation:

$$y = \mathbf{w} \cdot \mathbf{x} = \sum_{i,j} w_{i,j} x_{i,j}. \tag{1}$$

The scalars $w_{i,j}$ correspond to the (synaptic) weights of a linear filter and are sometimes referred to as "the linear receptive field". These weights can be estimated empirically by presenting white noise as an input to the neuron while recording its response.

A good parametrization of such linear receptive field for the simple cells found in the primary visual cortex [36] is the Gabor function [43], which is given by the following equation:

$$w_{ij} = \exp\left(-\frac{(u^2 + \gamma^2 v^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} u + \phi\right) \tag{2}$$

$$\text{s.t. } u = i\cos\theta + j\sin\theta \text{ and } v = -i\sin\theta + j\cos\theta. \tag{3}$$

The five parameters, *i.e.*, the orientation θ, the aspect ratio γ, the effective width σ, the phase ϕ and the wavelength λ, determine the properties of the spatial receptive field of the corresponding model simple cell. Fig. 1 shows examples of model simple cells varying in orientation, spatial frequency and phase. The Gabor function thus describes a process that spans from the visual input falling onto the retina on the one end, to the output activity of a model simple cell in the primary visual cortex on the other end, including intermediate stages such as processing by the lateral geniculate nucleus (LGN).

Other parametrizations have been proposed for simple cells. These include Gaussian derivatives which have been shown to provide an excellent fit to cortical cells receptive fields both in the spatial [110] and spatio-temporal domain [111]. They were used in one of the first models of pre-attentive texture discrimination using early vision mechanisms [54].

However, biological neurons also behave in nonlinear ways; *e.g.*, their output tends to saturate as their input grows stronger, instead of increasing indefinitely as a linear input-output function would predict. Thus, cortical filter models

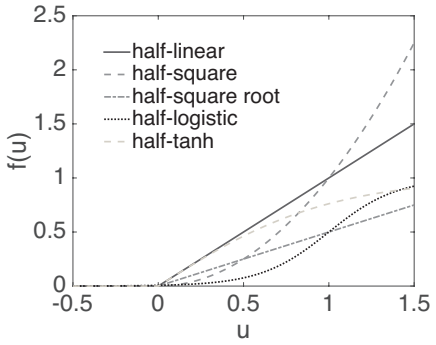

Figure 2: **Common nonlinear transfer functions used in cortical filter models of the primary visual cortex**. See text for details.

always include a nonlinear transfer function following the linear part, explaining why they are also referred to as linear-nonlinear (LN) models. The above input-output function of a model neuron then becomes:

$$y = f(\mathbf{w} \cdot \mathbf{x}), \tag{4}$$

where f is a nonlinear transfer function. Popular choices for the function f include (half-) linear rectification functions, exponential functions (square and square root), or the logistic and the hyperbolic functions (see Fig. 2). The LN model has been shown to account for a host of experimental data [79] and it has been shown that in many cases, biophysically more realistic models of neurons (which include a spike generation process) can be reduced to a simple LN model [69].

1.2 Divisive normalization

An extension of the LN model includes the addition of a normalization stage:

$$y = \frac{f(\mathbf{w} \cdot \mathbf{x})}{k + \sum_{j \in J} g(\mathbf{v} \cdot \mathbf{u^j})}, \tag{5}$$

where $k > 0$ is a constant to avoid division by zero. The pool of units used for normalization, indexed by $j \in J$, may correspond to the same (or another) set of input units as in the *tuning circuits* [47] or another set of output units as in the divisive normalization model [33]. In addition, the spatial extension of the set of units (whether input or output) may be limited to the (classical) receptive field of the unit, or possibly extend beyond to account for extra-classical receptive field effects [89] (see Fig. 3 for a common implementation of divisive normalization in the primary visual cortex). Normalization circuits were originally proposed to explain the contrast response of cells in the primary

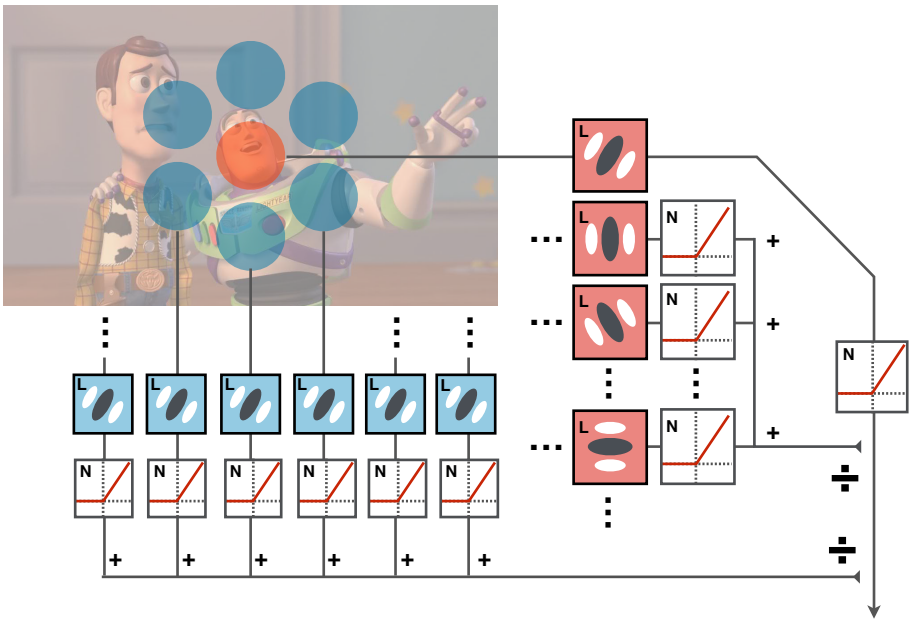

Figure 3: **Divisive normalization in the primary visual cortex**. The inhibited unit (receptive field shown as a red circle) gets inhibited by other units which share the same tuning preference but with receptive fields located outside its receptive field (blue circles; this region is called the "extra-classical receptive field" or "surround region"). This unit also gets inhibited by other units within the same cortical hypercolumn (in red). "L" and "N" respectively denote the linear and nonlinear part of the cortical LN model.

visual cortex [33] and are now thought to operate throughout the visual system, and in many other sensory modalities and brain regions (see [6] for a recent review).

For instance, in the HMAX model [80, 93] (see also Fig. 8 and discussion later), two types of operations are being assumed: a bell-shape tuning operation for simple cells and a max-like operation (or "soft-max") at the level of position and scale-tolerant complex cells [80]. Interestingly, both operations can be approximated as specific instances of the more general Eq. 5:

$$y = \frac{\sum_{i,j} w_{i,j} \; x_{i,j}^p}{k + \left(\sum_{i,j} x_{i,j}^q \right)^r}, \tag{6}$$

where p, q and r represent static nonlinearities in the underlying neural circuit.

An extra sigmoid transfer function on the output $g(y) = 1/(1 + \exp^{\alpha(y-\beta)})$ controls the sharpness of the unit response. By adjusting these nonlinearities, Eq. 6 can approximate well either a max operation or a tuning function (see [47] for details).

1.3 LN cascade

More sophisticated computations in visual cortex can also be captured by cascading several LN models into a single pipeline, where several stages of processing take place in succession: the output of one stage, described by the output of one LN model, can be integrated into the input of the following LN model describing the next stage. In the notations used above, it means that the input $x_{i,j}$ to a model cell does not represent a direct input from the visual field anymore; instead, each position i, j in that cell's receptive field correspond to an actual output y from a model unit from the previous stage. The corresponding linear weights $w_{i,j}$ then constitute a "generalized receptive field".

One example include Hubel & Wiesel's model of position invariance at the level of complex cells. Such invariance is obtained by locally pooling over simple cells with the same preferred orientation from the afferent layer. Another instance of the LN cascade was used to account for complex cells' invariance to contrast reversal. Unlike simple cells that are sensitive to the polarity of a stimulus presented (*e.g.*, white bar on a black background as opposed to a black on white background), complex cells exhibit a response which is largely invariant to such change. One circuit that has been proposed to explain this type of invariance is the energy model [1, 21, 88]. In the proposed circuit, the activity of a set of simple cells (corresponding to a first LN processing stage parametrized by a Gabor function) with the same preferred selectivity for orientation and spatial frequency but different selectivity for phases (corresponding to different preferred contrast polarity) are squared, then summed (sometimes followed by a square root nonlinearity) by a second LN stage.

$$y = \sqrt{\sum_{\phi} f\left(x_{\phi}\right)^2}, \tag{7}$$

Eq. 7 guarantees the result to be invariant to the reversal of the image contrast as contrast dependence is modeled by the phase parameter in the Gabor function. In signal processing theory, this computation is called the energy function as it is equivalent to a local measure of the amplitude spectrum of the image. The energy is taken over two phases, which are said to be in quadrature, if the Gabor functions are followed by a full rectification (more realistic circuits include more afferent subunits [2, 87]). Thus, the invariance to contrast reversal of complex cells is also known as an invariance to phase, by analogy with the phase parameter of the Gabor model.

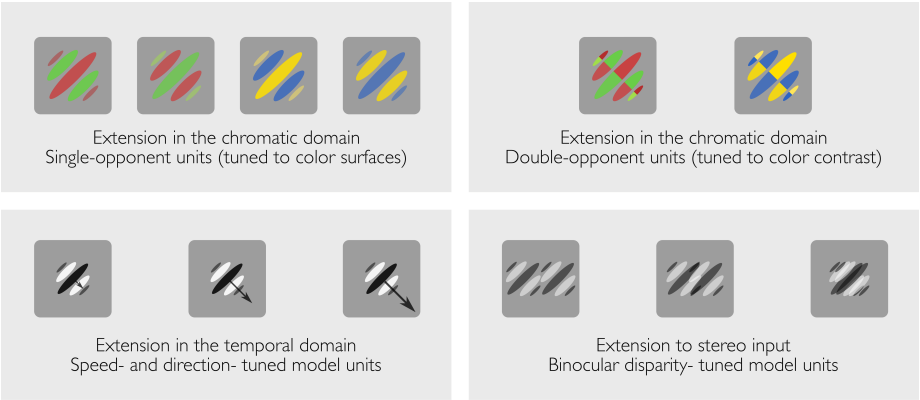

Figure 4: **Extension of the Gabor function to multiple visual modalities**. The notion of spatial filtering highlighted for the processing of two-dimensional shape information can be extended to the color, disparity and space-time (*i.e.*, motion) domains.

2 Cortical filter models across visual cues

As discussed above, the Gabor function given in Eq. 3 provides a good description of the response of simple cells' receptive fields that are characterized by a preferred orientation, spatial frequency, phase and bandwidth tuning. However, the conventional Gabor function only explains a limited range of the selectivity of cells observed in the primary visual cortex, namely the processing of two-dimensional shape information and local contrast. As we will show next, it is possible to generalize this simple cortical filter model to account for the processing of additional visual cues including color, motion, and binocular disparity.

Beyond simple cells, there exist complex cells tuned for motion, disparity and color. However, in order to understand how complex cells should be wired across visual channels, it is important to consider their computational roles within the context of different visual functions: beyond position invariance which was the original focus in models of complex cells' visual processing [36], one needs to consider the cue-specific invariances (and selectivities) that are relevant to a particular visual channel. Although the answer is often specific to each cue, we will strive to highlight common theoretical principles whenever possible.

2.1 Color processing

The standard Gabor function is a local function of the image contrast. Studies of color processing have shown that three types of cones in the retina, that are selective for long (L), medium (M) and short (S) wavelengths, project to the LGN via ganglion cells. In the LGN, the visual input is reorganized into opponent color channels, which are also found in the primary visual cortex: Red (R) versus green (G) and blue (B) versus yellow (Y). The existence of additional

channels such as a red versus cyan channel is also debated [8]. These channels can be traced back to inputs from individual cone types [95] (either L versus M, or S versus L+M, respectively). The conventional Gabor function, defined by its shape parameters, can be extended to the chromatic domain in order to account for the color-sensitive receptive fields of primary visual cortex simple cells [112].

There exist two functional classes of color-sensitive cells: single- and double-opponent receptive fields [37, 64, 95]. Single-opponent receptive fields exhibit a center-surround configuration with excitatory center and inhibitory surround corresponding to one of the following pairs: R+G- (red ON, green OFF, meaning the cell is driven by a red light increment in the center and inhibited by an increment of green in the surround), G+R-, B+Y-, and Y+B-. Electrophysiological studies have shown that such cells can be modeled well by considering a standard Gabor function, and using the positive component of the function for the ON component of the receptive field, and its negative component for the OFF component [42, 95]. Note that this yields only a weakly orientation-selective receptive field. In a computer model, the Gabor function can be defined across the L, M and S channels of an LMS image (or more simply across the R, G and B channels of an RGB image for a computer vision application [112]).

Beside single-opponent cells, one can find double-opponent receptive fields, which, in addition to exhibiting chromatic opponency, also exhibit spatial opponency. Zhang *et al.* [112] have shown that this type of receptive field organization can be derived by cascading the output of the single-opponency (LN + normalization) stage with an additional LN stage (Gabor function + half-wave rectification, see Fig. 5).

Note that although we described how to extend a Gabor receptive field to account for color opponent processing, color-selective complex cells require additional work. The chromatic analogue of the invariance to contrast reversal as observed in the grayscale domain is an invariance to chromatic contrast reversal, or, equivalently, to the phase of a color-opponent equiluminant grating [40, 41, 42]. To achieve such an invariance, one may consider an energy model atop an LN stage based on a chromatic Gabor receptive field to combine the output of a model cell selective for a certain opponent pair with that of a model cell selective for the reverse pair. For example, a red-green double opponent complex cell can be created by combining the activities of double-opponent simple cells R+G- and G+R- with the energy model. In addition, the resulting model complex cell can also be made invariant to the position of its preferred stimulus within its receptive field, just like a conventional contrast-sensitive complex cell, by using a max-pooling operation as discussed before. In the end, the resulting color-responsive complex cell responds most to a chromatic edge or bar of the preferred orientation and opponent color pair anywhere in its receptive field.

2.2 Binocular disparity processing

The monocular Gabor function can be extended to binocular visual inputs [46]. By allowing the Gabor function associated with either eye to have independent

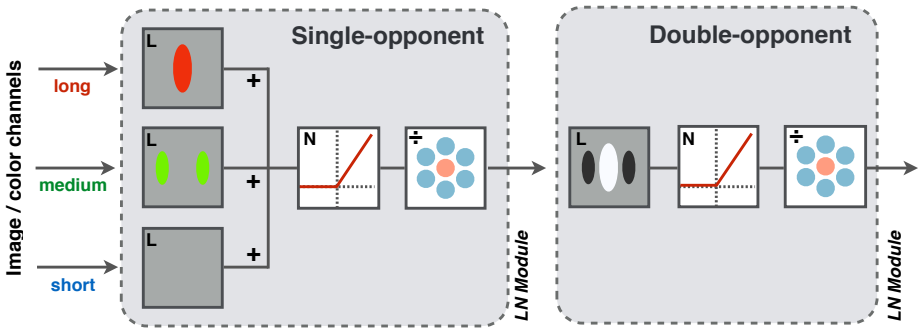

Figure 5: **The circuitry of single-opponent and double-opponent color-responsive units.** Single-opponent linear receptive fields in the primary visual cortex combine chromatically-opponent subunits (here in green and red for a R+G- unit), whereas double-opponent subunits include an additional step with a spatially-opponent, orientation-selective linear receptive field. Double-opponent processing is a good example of an LN cascade, where each LN module (boxes shaded in gray) is made of a canonical sequence of computations: a linear operation (L), then a nonlinear transfer function (N), and finally divisive normalization among several units (÷).

parameters, we introduce two new dimensions to the binocular filter thus created and that was described in section 1. Phase disparity is defined as the phase difference between the monocular inputs from each eye, and position disparity as the offset between the locations of either monocular receptive field within the visual field. A model binocular unit then linearly combines the output of each of its two afferents (given by a Gabor function for the left eye and another one for the right eye), followed the energy model described by Eq. 7; by extension, this particular model of binocular disparity processing is also called the energy model [68, 66]. By definition, such a cell exhibits a preference for a given phase and position disparity, and a population of such cells is able to represent all disparities across the visual field [76]. Disparity is a good representation to have as any object seen from a stereoscopic sensory device creates a specific pattern of disparity that is related to its viewing depth and appearance; thus, by leveraging disparity, depth can be recovered (more specifically, it is inversely related to disparity).

Given certain phase and position disparities, the notion of complex cell can be directly derived from the monocular case, by taking the energy of the cells with the same tuning preferences (both disparity, orientation and spatial frequency) and with their phases in quadrature [67, 66] (note that the quadrature is defined across afferents, which have binocular receptive fields, and not between each eye). Such a unit is selective for certain phase and position disparities while being phase-invariant; invariance in position can still be achieved through max-pooling. Note that some models also pool over orientations and/or spatial frequencies in order to reduce noise and make sure the population will peak

at veridical disparities [22]; such cells are then robust to changes to the power spectrum of the input stimulus (if pooling over both orientation and spatial frequency). Complex cells tuned to binocular disparity, invariant to reversal of contrast, were found in the primary visual cortex [67].

2.3 Motion processing

The static Gabor function can be extended to the space-time domain by introducing a time-dependent phase term that makes the periodic part of the Gabor function drift over time within the unit receptive field [96, 13, 39, 3] (note that some models also move the location of the receptive field itself over time; however, no evidence of such mechanism has been found in neurophysiology so far). This generalizes the idea of sampling from the power spectrum of a static image by a population of simple cells to that of sampling from the power spectrum of a image sequence over time [1, 3]. The resulting cells are selective for spatial frequency, orientation, and temporal frequency [1, 63, 3].

As before, phase-invariant complex cells can be obtained using the energy model by combining the output activities of a pair of simple cells with the same preferred location and tuning preferences, except for a phase difference of 90 degrees. Such a population of cells implicitly codes for local velocity (speed and direction) in a manner invariant to the textural content of the moving element [39]. The reason for this is that the movie of any visual element translating at a uniform velocity has a planar power spectrum [107, 106, 96]; changes in texture merely redistributes its power spectrum within that plane.

Therefore, model cells can be designed that are truly speed and direction selective and locally invariant to texture, which help to alleviate the aperture problem[1], by pooling together simple cells whose spectral receptive fields lie on the appropriate plane [96, 65]. By also pooling over nearby locations, such motion-sensitive complex cells gain local position invariance. Evidence for such cells has been found in MT [96, 65], therefore this is not strictly speaking a computational model of the primary visual cortex.

3 Completing the hierarchy: models of the visual cortex

3.1 Hubel & Wiesel model

Hubel & Wiesel [36] provided the first qualitative description of the receptive field (RF) organization of neurons in the primary visual cortex. As mentioned in Section 1, they described two function classes of neurons: the simple and complex cells. Simple cells respond best to oriented stimuli (*e.g.*, bars, edges,

[1]The "aperture problem" reflects the inherent ambiguity associated with the direction of motion of a moving stimulus within the receptive fields (a small aperture) of neurons in early visual areas. Because of its limited receptive field, a motion-selective neuron will often produce identical responses for stimuli that vary greatly in their shape, speed and orientation.

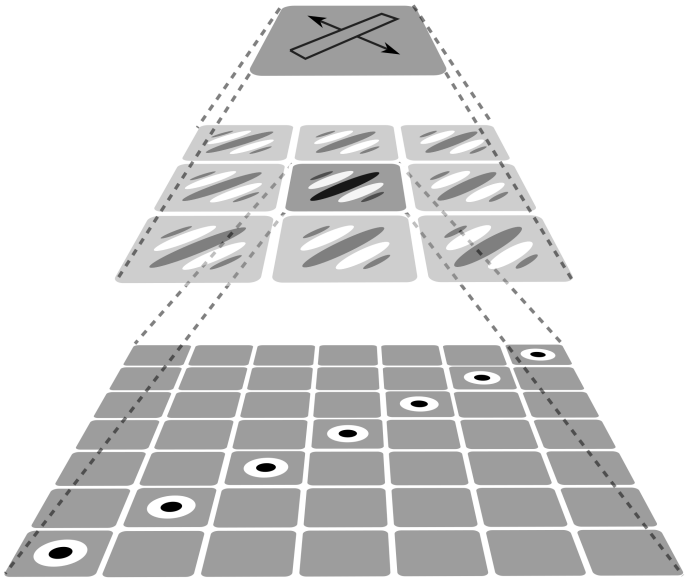

Figure 6: **Hubel & Wiesel model**. A simple unit (middle layer) pools over afferent units with center-surround receptive fields (bottom layer) aligned along a preferred axis of orientation. At the next stage (upper layer), a complex unit pools over afferent simple units with the same preferred orientation and spatial frequency within a small spatial neighborhood. Thus, the complex unit shown here is tolerant to local shifts of the preferred stimulus within its receptive field. A more complete model would also include pooling over simple cells tuned to slightly different spatial frequencies and phases [86, 7] – consistently with the observed broadening in frequency bandwidth [15] and tolerance to contrast reversal found in complex cells.

gratings) at one particular orientation, position and phase (*i.e.*, a light bar on a black background or a dark bar on a light background) within their relatively small receptive fields (typically a fraction of a degree up to about one degree of visual angle in the monkey). Complex cells, on the other end, while also selective for orientation, tend to have larger receptive fields (about twice as large) and exhibit some tolerance with respect to the exact position of the stimulus within their receptive fields. They are also invariant to contrast reversal, *i.e.*, the same cell responds to a white bar on a black background or the opposite.

Fig. 6 illustrates a plausible neural circuit proposed in [36] to explain the receptive field organization of these two functional classes of cells. Simple cell-like receptive fields can be obtained by pooling the activity of a small set of cells tuned to spots of lights with a center-surround organization (as observed in ganglion cells in the LGN and layer IV of the striate cortex) aligned along a preferred axis of orientation (Fig. 6, bottom layer). At the next stage, position tolerance at the complex cell level, can be obtained by pooling over afferent sim-

ple cells (from the level below) with the same preferred (horizontal) orientation but slightly different positions (Fig. 6, middle layer).

Today, nearly half a century after Hubel & Wiesel's initial proposal, the coarse circuitry underlying the organization of RFs in the primary visual cortex is relatively well established [2, 86, 7]. Using this circuit as a building block, numerous hierarchical models of the visual cortex have been proposed (see [90] for a recent review) and used to demonstrate the ability of this type of architectures to be invariant to increasingly challenging transformations in the visual input (*e.g.*, changes in the viewing angle of an object), while remaining selective for relevant aspects of it (*e.g.*, the identity of said object), a quandary also known as the invariance-selectivity trade-off.

3.2 Hierarchical models: formalism

Hierarchical models of the visual system come in many different forms: they differ primarily in terms of their specific wiring and corresponding parametrizations as well as the mathematical operations that they use. However, all these computational models exhibit a common underlying architecture corresponding to multiple cascaded stages of processing such as the one shown on Fig. 7. Units at any stage $k+1$ pool selectively over afferent units from the previous stage k over a local neighborhood (shown in pink). In general, pooling may occur over multiple dimensions of the afferent units (*e.g.*, position, scale, orientation, *etc.*). Pooling over multiple locations (as shown in stages k or $k+1$ on Fig. 7) leads to an increase in the receptive field size of the units at the next stage (compare the receptive field size of a unit at stage k shown in red with that of a unit at a higher stage $k+1$.

For instance, a computational instantiation of the Hubel & Wiesel hierarchical model of the primary visual cortex corresponds to three processing stages. Simple units in layer $k=1$ (highlighted in pink in Fig. 7) receive their inputs from center-surround cells in LGN sensitive to light increment (ON) or light decrement (OFF) in the previous layer $k=0$ (in red). Complex cells in layer $k=2$ pool over afferent simple cells at the same orientation over a local neighborhood (shown on Fig. 7 in purple is a 4×4 neighborhood). These types of circuits have yielded several models of the primary visual cortex that have focused on explaining in reasonably good biophysical details the tuning properties of individual cells (*e.g.*, orientation, motion or binocular disparity, see [48] for a review).

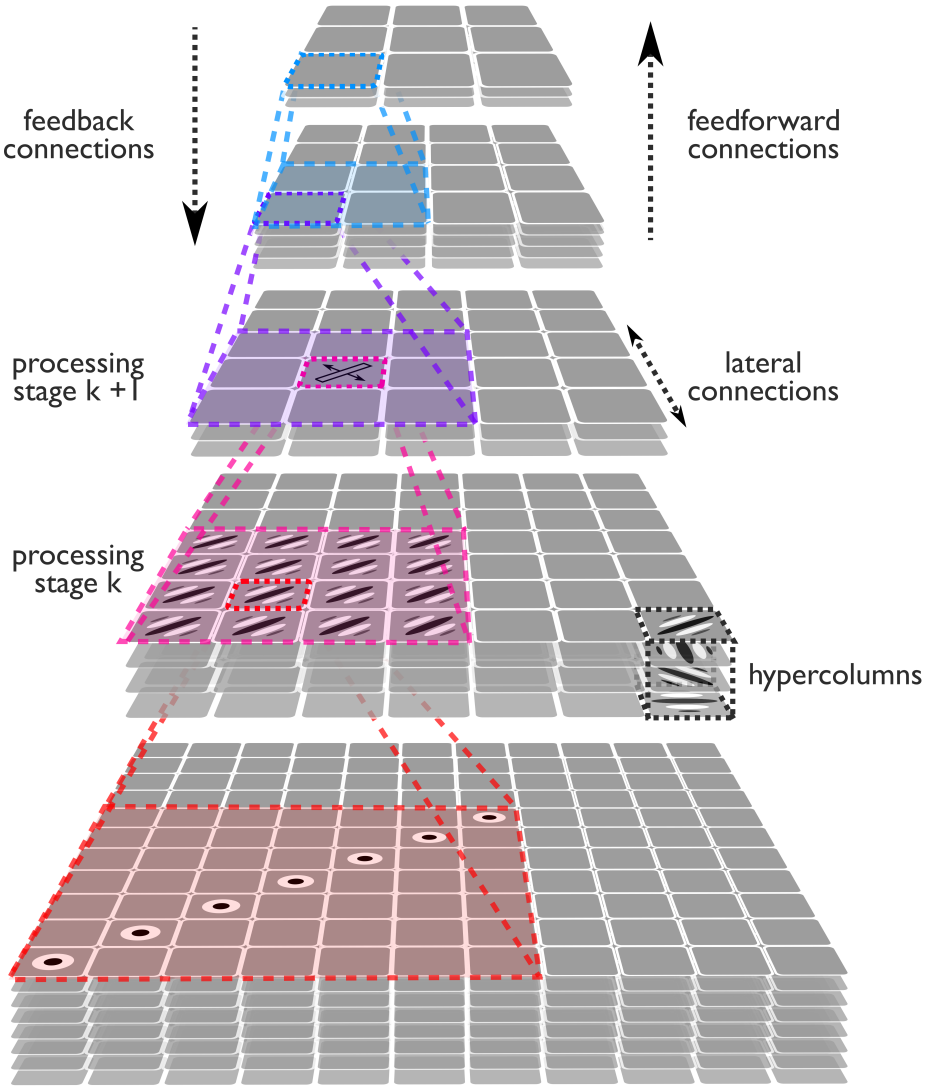

Figure 7: **Hierarchical models of the visual system.** They are characterized by multiple stages of processing whereby units in one stage (shown as squares) pool the response of units from the previous stage (colored projections). Individual stages, also called layers (shown as gray stacks), contain multiple *feature maps* organized in terms of both space and scale. An hyper-column contains all possible features from all feature maps for that location. Hence each stage can be thought of as containing hyper-columns replicated at all positions and scales. In the most general case, hierarchical models allow for communication both ways between any two consecutive stages (feedforward and feedback connections), as well as between units part of the same stage (lateral connections). The building block of many successful hierarchical models, *i.e.*, the circuit proposed by Hubel & Wiesel decades ago, is embedded as a particular example and shown in red (center-surround stage), pink (simple cell stage), and purple (complex cell stage).

In recent years, because of the increasing amount of computing power available, the scale of models of visual processing has increased with models now encompassing large portions of the visual field and entire streams of visual processing (see [91] for a review). Alternating between multiple layers of simple units and complex units leads to an architecture that is able to achieve a difficult trade-off between selectivity and invariance: along the hierarchy, units become tuned to features of increasing complexity (*e.g.*, from single oriented bars, to combinations of oriented bars to form corners and features of intermediate complexities) by combining afferents (complex units) with different selectivities (*e.g.*, units tuned to edges at different orientations). Conversely, at each "complex unit" stage, complex units become increasingly invariant to two-dimensional transformations (position and scale) by combining afferents (simple units) with the same selectivity (*e.g.*, a vertical bar) but slightly different positions and scales.

While recent work has suggested that 'simple' and 'complex' cells may represent the ends of a continuum instead of two discrete classes of neurons (see [83] for a discussion), this dichotomy is probably not critical for hierarchical models of the visual system. Indeed, recent models do not distinguish between simple and complex cell pooling [70].

Units in hierarchical models of the visual cortex are typically organized in columns and/or feature maps. An hyper-column (shown in blue in Fig. 7) corresponds to a population of units tuned to a basic set of features (*e.g.*, units spanning the full range of possible orientations or directions of motion, *etc.*)[2] in models of the primary visual cortex [see 36]. These hyper-columns are then replicated at all positions in the visual field and multiple scales. An alternative perspective is to think of processing stages in terms of feature maps. Typically, maps correspond to retinotopically organized population of units tuned to the same feature (*e.g.*, specific motion direction, orientation, binocular disparity, *etc.*) but at multiple positions (tiling the visual space) and/or multiple scales.

The first instance of such a columnar model was indeed proposed by Hubel & Wiesel [36] to explain orientation tuning and ocular dominance in the primary visual cortex, and was named the ice cube model. While more complex models of columnar organization have been proposed in recent years (*e.g.*, to account for pinwheel centers), hierarchical models of the visual system follow the inspiration of the ice cube model for its simplicity of implementation. Thus, the set of all units within a stage typically exhibits this kind of dual organization both in terms of the visual field which they tile with their receptive fields, and in terms of their selectivities, which can be thought to span all the possible values at every location in the visual field.

In addition to the feedforward (bottom-up) connections, which correspond to projections from processing stage k to $k* > k$, units can also be connected via lateral (horizontal) connections (both short-range connections within an hyper-column and long range between hyper-columns at different retinal locations) or feedback (top-down) connections from processing stage k to $k* < k$.

[2] A full model would also include eye dominance.

3.3 Models of object recognition

Historically, most hierarchical models that have been proposed have focused on the processing of two-dimensional shape information in the ventral stream of visual cortex, which follows a hierarchy of brain stages, starting from the retina, through the LGN in the thalamus to primary visual cortex (primary visual cortex, or striate cortex) and extra-striate visual areas, secondary visual cortex (V2), quaternary visual cortex (V4) and the inferotemporal cortex (IT). In turn, IT provides a major source of input to prefrontal cortex (PFC) involved in linking perception to memory and action (see [16] for a recent review).

As one progresses along the ventral stream visual hierarchy, neurons become selective for increasingly complex stimuli – from simple oriented bars and edges in early visual areas to moderately complex features in intermediate areas (such as combinations of orientations) and complex objects and faces in higher visual areas such as IT. In parallel to this increase in the complexity of the preferred stimulus, the invariance properties of neurons also increase with neurons gradually becoming more and more tolerant with respect to the exact position and scale of the stimulus within their receptive fields. As a result of this increase in invariance properties, the receptive field size of neurons increases, from about one degree or less in the primary visual cortex to several degrees in IT.

Explaining the selectivity and invariance properties of the ventral stream of the visual cortex has been one of the driving forces behind the development of hierarchical models of object recognition (see [90] for review). These models have a long history: the initial idea was proposed by Marko & Giebel with their homogeneous multi-layered architecture [56] and was later used in several visual architectures including Fukushima's *Neocognitron* [25], convolutional networks [50] and other models of object recognition [104, 61, 80, 108, 102, 93, 59]. Over the years, these hierarchical models were shown to perform well for the categorization of multiple object categories (see [94] for a review).

The HMAX model [80, 93] shown in Figure 8 constitutes a representative example of feedforward hierarchical models. It combines mechanisms for building up invariance and selectivity through the hierarchy, inspired by the *Neocognitron* with view-based theories of 3D object recognition [81]. HMAX attempts to mimic the main information processing stages across the entire ventral stream visual pathway and bridges the gap between multiple levels of understanding [94]. This system-level model seems consistent with physiological data in non-human primates in different cortical areas of the ventral visual pathway [92], as well as human behavioral data during rapid categorization tasks with natural images [93, 10] (but see also [27, 4, 109, 44]).

In recent years, a number of HMAX extensions have been proposed. Most of them have focused on the learning of visual representations in intermediate stages of the model. One prominent example includes the work by Masquelier and colleagues who incorporated biologically-plausible learning mechanisms in the HMAX based on temporal continuity in video sequences [58], evolutionary algorithms [28], as well as spike-timing dependent-based learning rules [59, 45].

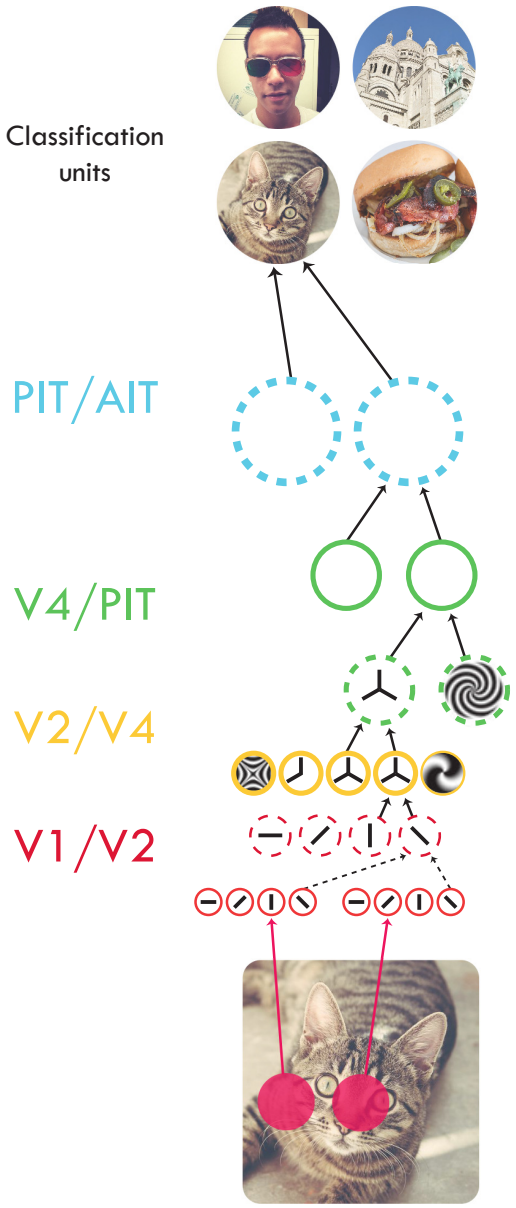

Figure 8: **Sketch of the Hmax hierarchical model of visual processing:** Acronyms: V1, V2 and V4 correspond to primary, secondary and quaternary visual areas, PIT and AIT to posterior and anterior inferotemporal areas, respectively. Tentative mapping to neurophysiology is shown in color, some areas of the parietal cortex and dorsal streams are not shown. The model relies on two types of computations: a max-like operation (shown in dash circles) over similar features at different position and scale to gradually build tolerance to position and scale, and a bell-shape tuning operation (shown in plain circles) over multiple features to increase the complexity of the underlying representation, see [91] and text for details. Since it was originally developed, the model has been able to explain a number of new experimental data. This includes data that were not used to derive or fit model parameters. The model seems to be qualitatively and quantitatively consistent with (and in some cases actually predicts) several properties of subpopulations of cells in the primary visual cortex, V4, IT, and PFC as well as fMRI and psychophysical data.

The models of object recognition described above use (Hebbian-like) *unsupervised* learning rules: they learn commonly-occurring visual features from natural images irrespective of their diagnosticity in object categorization. These learning rules seem consistent with ITC recordings that have shown that the learning of position and scale invariance, for instance, is driven by the subject's visual experience [51, 52] and is unaffected by reward signals [53].

However, a class of neural networks called deep learning architectures have recently brought about a small revolution in machine learning by becoming the new state-of-the-art on a variety of categorization tasks ranging from speech, music, text, genomes and images (see [49] for an up-to-date review). They differ in two ways from more traditional hierarchical models of the visual cortex such as the aforementioned HMAX and *Neocognitron*. First, learning across processing stages is fully supervised and uses the back-propagation algorithm (see [49] for a history), which propagates an error signal from upper-level (categorization) layers towards lower-level (perceptual) ones. Thus, only visual features that are diagnostic for the trained categorization tasks will be learned.

Second, deep learning architectures do not try to imitate biology as well as older hierarchical models of the visual cortex, which are constrained to match neuroscience data on a wide range of parameters (receptive field sizes, invariance and other tuning properties, number of layers). For instance, state-of-the-art deep learning architectures incorporate many more layers (over 20 layers [100, 32]) in comparison to hierarchical models of the visual cortex (*e.g.*, 7 layers for the HMAX). They possibly incorporate entire ensembles of deep networks for a given categorization task [100, 32]; improved training methods and accuracy have resulted in even deeper networks that can implement more complex classification functions. This, in turn, comes at the cost of sample complexity: the number of samples required for proper training increases with the number of parameters to be fitted. Not surprisingly, significant efforts have thus been recently dedicated to building ever-growing large-scale annotated image and video datasets (the ImageNet Large Scale Visual Recognition Challenge [84] contains >1M images and 1,000 categories), enabling the training of increasingly large networks (compare with the 2010 PASCAL VOC challenge [18] with <20,000 images and 20 categories).

Perhaps surprisingly, despite the absence of neuroscience constraints on modern deep learning architectures, recent work has shown that these architectures are better able to explain ventral stream neural data [109, 4, 44, 30]. In addition, these networks outperform all other models by a large margin [4] and are starting to match human level of accuracy for difficult object categorization tasks [32].

3.4 Models across visual cues

The effectiveness of hierarchical models of two-dimensional shape processing and object recognition has recently led to considerable interest in building hierarchical extensions to multiple visual cues beyond the early processing models described in Section 2. The main idea in these models is to reuse basic compu-

tational building blocks (such as the ones described in Section 1 and 2) across several processing stages. Moreover, the ever-increasing trove of electrophysiology data in mid-level visual areas now makes it possible to effectively constrain the space of all possible models.

Going beyond two-dimensional shape processing, several hierarchical models of motion processing have been proposed. For instance, computational models composed of the core operations described in Sections 1 and 2 have been shown to be able to reproduce the selectivity of motion-selective neurons in the dorsal stream of the visual cortex to complex moving stimuli such as drifting plaids [96, 85] and continuous deformations [62]. Closely related models of the ventral and dorsal streams for the processing of form and motion, respectively, were used to model the brain mechanisms underlying action recognition [29].

Building on models of the dorsal stream of the visual cortex [96, 29, 85, 62], a computer vision system was shown to perform well and, at the time, compete with state-of-the-art computer vision systems for the recognition of actions [39]. The approach was later extended to the automated monitoring and analysis of rodents in their home-cage with accuracy on par with that of trained human annotators for a repertoire of about a dozen behaviors [38].

More recently, an extension of this approach included speed-tuned units as found in MT [60, 71, 75] and yielded a system for the visual control of locomotory behavior that produced trajectories consistent with those produced by human participants when asked to reach a goal while avoiding obstacles in natural-looking environments (Barhomi *et al.* A data-driven approach to learning strategies for the visual control of navigation. Abstract presented at the Vision Science Society, 2014).

As for hierarchical models of color, the few efforts that have endeavored to go beyond one stage of processing pertain to computer vision (see [112] for an attempt to bridge this gap). These are very largely limited to solving specific tasks such as boundary detection in natural scenes and the representations they yield are *ad hoc* and cannot be compared against electrophysiology. However, work in progress from Zhang and colleagues suggests that a model consisting of the single- or double-opponent cells as described in Section 2.1 followed by the proper divisive normalization over an extended spatial neighborhood seems sufficient to account for psychophysics data of color constancy (Mély & Serre. A canonical circuit for visual contextual integration explains induction effects across visual modalities. Abstract presented at the Vision Science Society, 2015).

Regarding binocular disparity, our group has started to design a hierarchical model of disparity tuning [46] that builds on a population of model cells with linear receptive fields based on the binocular Gabor filters described in 2.2. Even though these units display varied selectivity to position disparity, phase disparity, orientation, spatial frequency, scale and phase, they are individually prone to incorrectly matching visually discordant inputs from either eye. To address this problem (see [77] for a formalization), we leveraged the divisive normalization circuit between units that prefer the same position disparity but opposite phase disparities in order to reduce sensitivity to false matches. We

further included an energy computation as well to implement local invariance to stimulus phase. As a result, units from this additional stage of the model tend to be much more selective to the correct binocular disparity.

4 Discussion and concluding remarks

4.1 Why hierarchies?

It has been postulated that the goal of the visual cortex is to achieve an optimal trade-off between selectivity and invariance via a hierarchy of processing stages whereby neurons at higher and higher levels exhibit an increasing degree of invariance to image transformations such as translations and scale changes [80, 94].

Now, why hierarchies? The answer – for models in the Hubel & Wiesel spirit – is that the hierarchy may provide a solution to the invariance-selectivity trade-off problem by decomposing a complex task such as invariant object recognition in a hierarchy of simpler ones (at each stage of processing). Hierarchical organization in cortex is not limited to the visual pathways, and thus a more general explanation may be needed. Interestingly, from the point of view of classical learning theory [74], there is no need for architectures with more than three layers. So, why hierarchies? There may be reasons of efficiency, such as the efficient use of computational resources. For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks [26].

There may also be the more fundamental issue of sample complexity, the number of training examples required for good generalization (see [94] for discussion). An obvious difference between the best classifiers derived from statistical learning theory and human learning is in fact the number of examples required in tasks such as object recognition. Statistical learning theory shows that the complexity of the hypothesis space sets the speed limit and the sample complexity for learning. If a task – like a visual recognition task – can be decomposed into low-complexity learning tasks for each layer of a hierarchical learning machine, then each layer may require only a small number of training examples. Neuroscience suggests that what humans can learn may be represented by hierarchies that are locally simple. Thus, our ability to learn from just a few examples, and its limitations, may be related to the hierarchical architecture of cortex.

4.2 Limitations

To date, most existing hierarchical models of visual processing both from the perspective of biological and machine vision are instances of feedforward models. These models have been useful to explore the power of fixed hierarchical organization as originally suggested by Hubel & Wiesel. These models assume that our core visual capabilities proceed through a cascade of hierarchically

organized areas along various streams of processing in the visual cortex with computations at each successive stage being largely feedforward [80, 16]. They have led, for instance, to algorithms that were at the time competitive with the best computer vision systems [94] and culminating with deep learning architectures that are bringing about a small revolution in artificial intelligence [49].

The limitations of these visual architectures, however, are becoming increasingly obvious. Not only top-down effects are key to normal, everyday vision, but back-projections are also likely to be a key part of what cortex is computing and how. Thus, a major question for modeling visual cortex revolves around the role of back-projections and the related fact that vision is more than categorization and requires interpreting and parsing visual scenes (as opposed to simply finding out whether a specific object is present in the visual scene or not). A human observer can essentially answer an infinite number of questions about an image (one could in fact imagine a Turing test of vision). Such image interpretation tasks have proven challenging for modern computer vision architectures [23, 31].

In addition, while the overall hierarchical organization of the visual cortex is now well established [19], the parallel between the anatomical and functional hierarchy is, however, looser than one might expect. While the trend is, from lower to higher visual areas, for neurons' receptive fields to become increasingly large and tuned to increasingly complex preferred stimuli, there remains a very broad distribution of tuning and receptive field sizes in all areas of the visual hierarchy. For instance, IT, which is commonly assumed to have solved the problem of invariant recognition [16], also contains neurons with relatively small receptive fields and tuned to relatively simple visual features such as simple orientations [14]. A close comparison of shape representation between primary visual cortex, V2 and V4 also demonstrated a complex pattern of shape selectivity with significant deviation from strict hierarchical organization with some cells in the primary visual cortex exhibiting more complex tuning than some cells in V4 [34]. Furthermore, beside the visual cortical hierarchy, there exist additional subcortical pathways (including cortico-thalamo-cortical loops). Hence, the anatomical hierarchy should be taken as an idealization and cannot be taken as a strict flowchart of visual information [35].

Another weakness shared by both larger-scale models of biological and machine vision are their reliance on a surprisingly limited number of computations, *viz.*, the linear-nonlinear (LN) modules we mentioned in Section 1 and divisive normalization under various forms. As a result, a potentially fruitful way to improve such hierarchical models would be to extend their repertoire to include new computations inspired by cutting-edge neurophysiology research on cortical microcircuits. Among the cortical operations yet untapped on a large scale by modeling efforts are dynamic or stochastic synapses (current models assume synaptic weights to be fixed and static after learning), heavily nonlinear computations in dendrites (current models only assume "weak" nonlinearity in their units, *viz.* linearity followed by a rectification, as opposed to more complex transformation), or synchrony between model neurons (though many researchers have discussed its potential use to tackle the "binding problem' between overlap-

ping, noisy visual representations [9, 103, 97, 24, 101, 98, 78]). The availability of large-scale architectures such as deep learning nets, combined with extensive human-annotated datasets, should make for an ideal testbed for any potential, cortically-inspired computational mechanism.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A.*, 2(2):284–299, 1985.

[2] J. M. Alonso and L. M. Martinez. Functional connectivity between simple cells and complex cells in cat striate cortex. *Nat. Neurosci.*, 1(5):395–403, 1998.

[3] D. Bradley and M. Goyal. Velocity computation in the primate visual system. *Nat. Rev. Neurosci.*, 9(9):686–695, 2008.

[4] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.*, 10(12):e1003963, Dec. 2014.

[5] M. Carandini. From circuits to behavior: a bridge too far? *Nat. Neurosci.*, 15(4):507–9, Apr. 2012.

[6] M. Carandini and D. Heeger. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.*, 13:51–62, Nov. 2012.

[7] X. Chen, F. Han, M.-m. M. Poo, and Y. Dan. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). *Proc. Natl. Acad. Sci. U. S. A.*, 104(48):19120–5, Nov. 2007.

[8] B. R. Conway. Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1). *J. Neurosci.*, 21(8):2768–83, Apr. 2001.

[9] F. Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *Proc Natl Acad Sci USA*, 81:4586–4590, 1984.

[10] S. M. Crouzet and T. Serre. What are the Visual Features Underlying Rapid Object Recognition? *Front. Psychol.*, 2:326, Jan. 2011.

[11] J. G. Daugman. Two-Dimensional Spectral Analysis of Cortical Receptive Field Profile. *Vision Res.*, 20:847–856, 1980.

[12] J. G. Daugman. Uncertainty Relation for resolution in Space, Spatial frequency, and Orientation Optimization by Two-dimensional Visual cortical Filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, 1985.

[13] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

[14] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.*, 4(8):2051–2062, 1984.

[15] R. L. DeValois, D. G. Albrecht, and L. G. Thorell. Spatial-frequency selectivity of cells in macaque visual cortex. *Vis. Res*, 22:545–559, 1982.

[16] J. J. Dicarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition ? *Neuron*, 73(3):415–434, 2012.

[17] R. J. Douglas and K. A. C. Martin. Mapping the matrix: the ways of neocortex. *Neuron*, 56(2):226–38, Oct. 2007.

[18] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[19] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. cortex*, 1:1–47, 1991.

[20] L. Fenno, O. Yizhar, and K. Deisseroth. The development and application of optogenetics. *Annu. Rev. Neurosci.*, 34:389–412, Jan. 2011.

[21] I. Finn and D. Ferster. Computational Diversity in Complex Cells of Cat Primary Visual Cortex. *J. Neurosci.*, 27(36):9638–9648, 2007.

[22] D. J. Fleet, H. Wagner, and D. J. Heeger. Neural encoding of binocular disparity: energy models, positionshifts and phase shifts. *Vis. Res.*, 36(12):1839–1857, 1996.

[23] F. Fleuret, T. Li, C. Dubout, E. K. Wampler, S. Yantis, and D. Geman. Comparing machines and humans on a visual categorization test. *Proc. Natl. Acad. Sci. U. S. A.*, 108(43):17621–5, Oct. 2011.

[24] P. Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn Sci*, 9(10):474–480, 2005.

[25] K. Fukushima. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, 13:826–834, 1983.

[26] D. Geman and A. Koloydenko. Invariant statistics and coding of natural microimages. *Proc. IEEE Work. Stat. Comput. Theor. Vis.*, 1999.

[27] M. Ghodrati, A. Farzmahdi, K. Rajaei, R. Ebrahimpour, and S.-M. Khaligh-Razavi. Feedforward object-vision models only tolerate small image variations compared to human. *Front. Comput. Neurosci.*, 8:74, Jan. 2014.

[28] M. Ghodrati, S.-M. Khaligh-Razavi, R. Ebrahimpour, K. Rajaei, and M. Pooyan. How can selection of biologically inspired features improve the performance of a robust object recognition model? *PLoS One*, 7(2):e32357, Jan. 2012.

[29] M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.*, 4(3):179–192, 2003.

[30] U. Guclu and M. A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.*, 35(27):10005–10014, July 2015.

[31] c. Gülçehre and Y. Bengio. Knowledge Matters: Importance of Prior Information for Optimization. *CoRR*, abs/1301.4, Jan. 2013.

[32] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Feb. 2015.

[33] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Vis. neurosci.*, 9(2):181–197, 1992.

[34] J. Hegdé and D. V. Essen. A comparative study of shape representation in macaque visual areas V2 and V4. *Cereb. Cortex*, 2(May), 2007.

[35] J. Hegdé and D. J. Felleman. Reappraising the functional implications of the primate visual anatomical hierarchy. *Neurosci.*, 13(5):416–21, Oct. 2007.

[36] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160:106–154, 1962.

[37] R. A. Humanski and H. R. Wilson. Spatial-frequency adaptation: evidence for a multiple-channel model of short-wavelength-sensitive-cone spatial vision. *Vis. res.*, 33(5-6):665–675, 1993.

[38] H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre. Automated home-cage behavioural phenotyping of mice. *Nat. Commun.*, 1(6):1–9, Sept. 2010.

[39] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A Biologically Inspired System for Action Recognition. *2007 IEEE 11th Int. Conf. Comput. Vis.*, pages 1–8, 2007.

[40] E. N. Johnson, M. J. Hawken, and R. Shapley. The spatial transformation of color in the primary visual cortex of the macaque monkey. *Nat. Neurosci.*, 4(4):409–16, Apr. 2001.

[41] E. N. Johnson, M. J. Hawken, and R. Shapley. Cone inputs in macaque primary visual cortex. *J. Neurophysiol.*, 91(6):2501–14, June 2004.

[42] E. N. Johnson, M. J. Hawken, and R. Shapley. The orientation selectivity of color-responsive neurons in macaque V1. *J. Neurosci.*, 28(32):8096–106, Aug. 2008.

[43] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol*, 58(6):1233–1258, 1987.

[44] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput. Biol.*, 10(11):e1003915, Nov. 2014.

[45] S. R. Kheradpisheh, M. Ganjtabesh, and T. Masquelier. Bio-inspired Unsupervised Learning of Visual Features Leads to Robust Invariant Object Recognition. *CoRR*, abs/1504.0, Apr. 2015.

[46] J. Kim, D. A. Mely, and T. Serre. A critical evaluation of computational mechanisms of binocular disparity.

[47] M. Kouh and T. Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural Comput.*, 20(6):1427–1451, 2008.

[48] M. S. Landy and J. A. Movshon. *Computational models of visual processing*. MIT Press, 1991.

[49] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE*, 86(11):2278–2324, Nov. 1998.

[51] N. Li and J. J. DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science (80-.).*, 321(5895):1502–1507, 2008.

[52] N. Li and J. J. DiCarlo. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron*, 67(6):1062–1075, 2010.

[53] N. Li and J. J. Dicarlo. Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *J. Neurosci.*, 32(19):6611–20, May 2012.

[54] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A.*, 7(5):923–932, 1990.

[55] S. Marcelja. Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.*, 70:1297–1300, 1980.

[56] H. Marko and H. Giebel. Recognition of handwritten characters with a system of homogeneous Layers. *Nachrichtentechnische Zeitschrift*, 23:455–459, 1970.

[57] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information.* W.H.Freeman & Co Ltd, San Francisco, 1982.

[58] T. Masquelier, T. Serre, and T. Poggio. Learning complex cell invariance from natural videos : A plausibility proof. Technical report, Massachusetts Institute of Technology, Cambridge MA, 2007.

[59] T. Masquelier and S. J. Thorpe. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol*, 3(2):e31, 2007.

[60] J. H. Maunsell and D. C. V. Essen. Functional properties of neurons in middle temporal visual area ofthe macaque monkey. II. Binocular interactions and sensitivity tobinocular disparity, 1983.

[61] B. W. Mel. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput.*, 9(4):777–804, 1997.

[62] P. Mineault, F. Khawaja, D. Butts, and C. Pack. Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proc. Natl. Acad. Sci.*, 109(16):E972–E980, 2012.

[63] J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome. The analysis of moving visual patterns. *Pattern Recognit. Mech.*, 1985.

[64] K. T. Mullen and M. A. Losada. The spatial tuning of color and luminance peripheral vision measured with notch filtered noise masking. *Vision Res.*, 39(4):721–31, Feb. 1999.

[65] S. Nishimoto and J. L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *J. Neurosci.*, 31(41):14551–64, Oct. 2011.

[66] I. Ohzawa. Mechanisms of stereoscopic vision: the disparity energy model. *Curr. Opin. Neurobiol.*, 8(4):509–15, Aug. 1998.

[67] I. Ohzawa, G. DeAngelis, and R. Feeman. Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.*, 77(6):2879–909, June 1997.

[68] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman. Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophysiol.*, 75(5):1779–805, May 1996.

[69] S. Ostojic and N. Brunel. From spiking neuron models to linear-nonlinear models. *PLoS Comput. Biol.*, 7(1):e1001056, Jan. 2011.

[70] R. C. OReilly, D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk. Recurrent Processing during Object Recognition. *Front. Psychol.*, 4(April):1–14, 2013.

[71] J. A. Perrone and A. Thiele. Speed skills: measuring the visual speed analyzing properties of primate MT neurons. *Nat. Neurosci.*, 4(5):526–32, May 2001.

[72] S. M. Plaza, L. K. Scheffer, and D. B. Chklovskii. Toward large-scale connectome reconstructions. *Curr. Opin. Neurobiol.*, 25:201–10, Apr. 2014.

[73] T. Poggio and T. Serre. Models of the visual cortex. *Scholarpedia*, 8(4):3516, 2013.

[74] T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Not. Am. Math. Soc.*, 50(5), 2003.

[75] N. J. Priebe, C. R. Cassanello, and S. G. Lisberger. The Neural Representation of Speed in Macaque Area MT/V5. *J. Neurosci.*, 23(13):5650–5661, July 2003.

[76] N. Qian. Computing Stereo Disparity and Motion with Known Binocular Cell Properties. *Neural Comput.*, 6(3):390–404, May 1994.

[77] J. C. Read and B. G. Cumming. Sensors for impossible stimuli may solve the stereo correspondence problem. *Nat. Neurosci.*, 10(10):1322–8, Oct. 2007.

[78] D. P. Reichert and T. Serre. Neuronal Synchrony in Complex-Valued Deep Networks. In *Int. Conf. Learn. Vis. Represent.*, Dec. 2014.

[79] F. Rieke, D. Warland, R. van Steveninck, W. Bialek, and R. van Steveninck. *Spikes*. The MIT Press, Cambridge, Massachusetts, 1997.

[80] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–25, Nov. 1999.

[81] M. Riesenhuber and T. Poggio. Models of object recognition. *Nat. Neurosci.*, 3:1199–204, Nov. 2000.

[82] D. L. Ringach. Haphazard wiring of simple receptive fields and orientation columns in visual cortex. *J. Neurophysiol.*, 92:468–476, 2004.

[83] D. L. Ringach. Mapping receptive fields in primary visual cortex. *J. Physiol.*, 558(3):717–28, Aug. 2004.

[84] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *CoRR*, abs/1409.0:43, Sept. 2014.

[85] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How MT cells analyze the motion of visual patterns. *Nat. Neurosci.*, 9(11):1421–31, Nov. 2006.

[86] N. C. Rust, O. Schwartz, J. A. Movshon, and E. P. Simoncelli. Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, 46(6):945–56, June 2005.

[87] T. M. Sanada and I. Ohzawa. Encoding of three-dimensional surface slant in cat visual areas 17 and 18. *J. Neurophysiol.*, 95(5):2768–86, May 2006.

[88] K. Sasaki and I. Ohzawa. Internal Spatial Organization of Receptive Fields of Complex Cells in the Early Visual Cortex. *J. Neurophysiol.*, 98(3):1194–1212, 2007.

[89] P. Series, J. Lorenceau, and Y. Frégnac. The silent surround of V1 receptive fields: theory and experiments. *J. Physiol.*, 97:453–474, 2003.

[90] T. Serre. Hierarchical Models of the Visual System, 2014.

[91] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio. A quantitative theory of immediate visual recognition. *Prog. Brain Res.*, 165:33, 2007.

[92] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio. A quantitative theory of immediate visual recognition. *Prog. Brain Res.*, 165(06):33–56, Jan. 2007.

[93] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U. S. A.*, 104(15):6424–6429, 2007.

[94] T. Serre and T. Poggio. A neuromorphic approach to computer vision. *Commun. ACM*, 53(10):54, Oct. 2010.

[95] R. Shapley and M. J. Hawken. Color in the Cortex: single- and double-opponent cells. *Vision Res.*, 51:701–717, Feb. 2011.

[96] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area MT. *Vision Res.*, 38(5):743–761, 1998.

[97] W. Singer and C. M. Gray. Visual feature integration and the temporal correlation hypothesis. *Ann. Rev. Neurosci.*, 18:555–586, 1995.

[98] G. B. Stanley. Reading and writing the neural code. *Nat. Neurosci.*, 16(3):259–63, Mar. 2013.

[99] I. H. Stevenson and K. P. Kording. How advances in neural recording affect data analysis. *Nat. Neurosci.*, 14(2):139–42, Feb. 2011.

[100] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *CoRR*, abs/1409.4, Sept. 2014.

[101] P. J. Uhlhaas, G. Pipa, B. Lima, L. Melloni, S. Neuenschwander, D. Nikolić, and W. Singer. Neural synchrony in cortical networks: history, concept and current status. *Front. Integr. Neurosci.*, 3:17, Jan. 2009.

[102] S. Ullman. Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn. Sci.*, 11(2):58–64, 2007.

[103] C. von der Malsburg. The Correlation Theory of Brain Function. In *Model. Neural Networks II*, pages 94–119. Springer-Verlag, e. domany, edition, 1994.

[104] G. Wallis and E. T. Rolls. A model of invariant recognition in the visual system. *Prog. Neurobiol.*, 51:167–194, 1997.

[105] A. B. Watson. Efficiency of a model human image code. *J. Opt. Soc. Am. A.*, 4(12):2401–2417, 1987.

[106] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *J. Opt. Soc. Am. A.*, 2(2):322–341, 1985.

[107] A. B. Watson, H. B. Barlow, and J. G. Robson. What does the eye see best? *Nature*, 302(5907):419–422, 1983.

[108] H. Wersing and E. Koerner. Learning optimized features for hierarchical models of invariant recognition. *Neural Comput.*, 15(7):1559–1588, 2003.

[109] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23):8619–24, June 2014.

[110] R. A. Young. The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spat. Vis.*, 2(4):273–293, 1987.

[111] R. A. Young and R. M. Lesperance. The Gaussian derivative model for spatial-temporal vision: II. Cortical data. *Spat. Vis.*, 14(3):321–389, 2001.

[112] J. Zhang, Y. Barhomi, and T. Serre. A new biologically inspired color image descriptor. In *Eur. Conf. Comput. Vis.*, volume 7576 LNCS, pages 312–324, 2012.