Abstract

Human and non-human primates excel at visual recognition tasks. In particular, the visual system of primates exhibits a strong degree of selectivity while at the same time being robust to changes in the input images. We have developed a quantitative theory to account for the computations performed by the feedforward path in the ventral stream of visual cortex. We here review recent predictions by a model instantiating the theory about physiological observations in higher visual areas. We also show that the model can perform recognition on datasets of complex natural images at a level comparable to psychophysical measurements on human observers during rapid categorization tasks. In sum, the evidence suggests that the theory may provide a framework to explain the first 100-150 milliseconds of visual object recognition. The model also constitutes a vivid example of how computational models can interact with experimental observations in order to advance our understanding of a complex phenomenon. We conclude by suggesting a number of open questions, predictions and specific experiments for visual physiology and psychophysics.


1 Introduction

The primate visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes. In particular, it can easily categorize images or parts of them, such as faces, and identify a specific one. This remarkable ability is of prime evolutionary importance since it allows us to distinguish friend from foe and identify food targets in complex crowded scenes. Despite the ease with which we see, visual recognition – one of the key issues addressed in computer vision – is quite difficult for computers. The problem of object recognition is even more difficult from the point of view of Neuroscience, since it involves several levels of understanding from the information processing or computational level to circuits and biophysical mechanisms. After decades of work in different brain areas ranging from the retina to higher cortical areas, the emerging picture of how cortex performs object recognition is in fact becoming too complex for any simple, qualitative “mental” model.

A quantitative, computational theory can provide a much needed framework for summarizing and integrating existing data and for planning, coordinating and interpreting new experiments. Models are powerful tools in basic research, integrating across several levels of analysis – from molecular, to synaptic, to cellular, to systems, to complex visual behavior. In this paper, we describe a quantitative theory of object recognition in primate visual cortex that (1) bridges several levels of understanding from biophysics to physiology and behavior and (2) achieves human level performance in rapid recognition of complex natural images. The theory is restricted to the feedforward path of the ventral stream and therefore to the first 100-150 ms or so of visual recognition; it does not describe top-down influences, though it should be, in principle, capable of incorporating them.

In contrast to other models that address the computations in any one given brain area (such as primary visual cortex) or attempt to explain a particular phenomenon (such as contrast adaptation or a specific visual illusion), we here describe a large-scale neurobiological model that attempts to describe the basic processes across multiple brain areas. One of the initial key ideas in this and many other models of visual processing [1–5] comes from the pioneering physiological studies and models of Hubel and Wiesel [6]. They established that neurons in primary visual cortex respond best if a bar of a specific orientation is presented within their receptive field. They suggested an elegant and simple model for this orientation preference by postulating that a neuron in primary visual cortex may receive its input from neurons in the thalamus whose receptive fields are aligned along the axis of orientation of the corresponding V1 neuron.

Beyond biologically-inspired algorithms [1, 7–9], i.e., systems only qualitatively constrained by the anatomy and physiology of the visual cortex, there have been very few neurobiologically plausible models [2–5, 10, 11] that try to address a generic, high-level computational function such as object recognition by summarizing and integrating a large body of data from different levels of understanding. What should a general theory of object recognition be able to explain? It should be constrained by different data from anatomy and physiology at different stages of the ventral stream and by the requirement of matching human performance in complex visual tasks such as object recognition. The theory we propose may well be incorrect. Yet it represents a set of claims and ideas that deserve to be either falsified or further developed and refined.

The scope of the current theory is limited to “immediate recognition”, i.e., to the first 100-150 ms of the flow of information in the ventral stream. This is behaviorally equivalent to considering “rapid categorization” tasks for which presentation times are fast and back-projections are likely to be inactive. For such tasks, presentation times do not allow sufficient time for eye movements or shifts of attention [12]. Furthermore, EEG studies [13] provide evidence that the human visual system is able to solve an object detection task – determining whether a natural scene contained an animal or not – within 150 ms. Extensive evidence shows that the responses
of inferior temporal (IT) cortex neurons begin 80-100 ms after onset of the visual stimulus [14]. Furthermore, the responses at the IT level are tuned to the stimulus essentially from response onset [15]. Recent data [16] show that the activity of small neuronal populations (around 100 randomly selected cells) in IT over very short time intervals (as small as 12.5 ms) from the onset of the neural response contains surprisingly accurate and robust information supporting visual object categorization and identification tasks. Finally, rapid detection tasks (e.g., animal vs. non-animal) [13] can be carried out without top-down attention [17]. Again, we wish to emphasize that none of this rules out the use of local feedback – which is in fact used by the circuits we propose for the main two operations postulated by the theory (see Section 2) – but suggests a hierarchical forward architecture as the core architecture underlying “immediate recognition”.

We start by presenting the theory in Section 2: we describe the architecture of a model implementing the theory, its two key operations and its learning stages. We briefly review the evidence about the agreement of the model with single cell recordings in visual cortical areas (V1, V2, V4) and describe in more detail how the final output of the model compares to the responses in IT cortex during a decoding task that attempts to identify or categorize objects (Section 3). In Section 4 we further extend the approach to natural images and show that the model performs surprisingly well in complex recognition tasks to the extent of competing with some of the best computer vision systems. As an ultimate and most stringent test of the theory, we show that the model predicts the level of performance of human observers on a rapid categorization task. The final Section 5 discusses the state of the theory, its limitations, a number of open questions, including critical experiments, and its extension to include top-down effects and cortical back-projections.

2 Quantitative framework for the ventral stream

2.1 Organization of the ventral stream of visual cortex

Object recognition in cortex is thought to be mediated by the ventral visual pathway [18]. Information from the retina is conveyed to the lateral geniculate nucleus in the thalamus and then to primary visual cortex, V1. Area V1 projects to visual areas V2 and V4, and V4 in turn projects to IT which is the last exclusively visual area along the ventral stream [19]. Based on physiological and lesion experiments in monkeys, IT has been postulated to play a central role in object recognition [20]. IT in turn is a major source of input to prefrontal cortex (PFC) which is involved in linking perception to memory and action [21].

Hubel and Wiesel [6] first described simple cells in V1 with small receptive fields that respond preferentially to oriented bars. Neurons along the ventral stream [2, 22, 23] show an increase in receptive field size as well as in the complexity of their preferred stimuli [24]. At the top of the ventral stream, IT cells are tuned to complex stimuli such as faces and other objects [14, 25, 26].

A hallmark of these IT cells is the robustness of their firing to stimulus transformations such as scale and position changes [2, 22, 23, 27]. In addition, as other studies have shown [2, 27–29], most neurons show specificity for a certain object view or lighting condition while other neurons were view-invariant in agreement with earlier predictions [30]. Whereas view-invariant recognition requires visual experience of the specific novel object, significant position and scale invariance seems to be immediately present in the view-tuned neurons [27] without the need of visual experience for views of the specific object at different positions and scales (see also [16]).

In summary, the accumulated evidence points to four, mostly accepted, properties of the feedforward path of the ventral stream architecture: a) A hierarchical build-up of invariances
first to position and scale and then to viewpoint and other transformations; b) An increasing selectivity, originating from inputs from previous layers and areas, with a parallel increase in both the size of the receptive fields and in the complexity of the optimal stimulus; c) A basic feedforward processing of information (for “immediate recognition” tasks); d) Plasticity and learning probably at all stages with a time scale that decreases from V1 to IT and PFC.

2.2 Architecture and model implementation

The physiological data summarized in the previous section, together with computational considerations on image invariances lead to a theory which summarizes and extends, several previously existing neurobiological models [2–6, 30] and biologically motivated computer vision approaches [1, 7, 8]. The theory maintains that:

1. One of the main functions of the ventral stream pathway is to achieve an exquisite trade-off between selectivity and invariance at the level of shape-tuned and invariant cells in IT from which many recognition tasks can be readily accomplished; the key computational issue in object recognition is to be able to finely discriminate between different objects or object classes while at the same time being tolerant to object transformations such as scaling, translation, illumination, viewpoint changes, change in context and clutter, non-rigid transformations (such as a change of facial expression) and, for the case of categorization, also to shape variations within a class.

2. The underlying architecture is hierarchical, with a series of stages in order to gradually increase invariance to object transformations and tuning to more specific and complex features;

3. There exist at least two main functional types of units, simple and complex, which represent the result of two main operations to achieve selectivity (S layer) and invariance (C layer). The two corresponding operations are a (bell-shaped) Gaussian-like tuning of the simple units and a max-like operation for invariance (to some degree) to position, scale and clutter of the complex units.

2.2.1 Two basic operations for selectivity and for invariance

The simple S units perform a TUNING operation over their afferents to build object-selectivity. The S units receive convergent inputs from retinotopically organized units tuned to different preferred stimuli and combine these subunits with a bell-shaped tuning function, thus increasing object selectivity and complexity of the preferred stimulus. Neurons with a Gaussian-like bell-shaped tuning are prevalent across cortex. For instance, simple cells in V1 exhibit a Gaussian tuning around their preferred orientation; cells in AIT are typically tuned around a particular view of their preferred object. From the computational point of view, Gaussian-like tuning profiles may be key in the generalization ability of cortex. Indeed networks that combine the activity of several units tuned with a Gaussian profile to different training examples have proved to be a powerful learning scheme [30].

The complex C units perform a MAX-like operation over their afferents to gain invariance to several object transformations. The complex C units receive convergent inputs from retinotopically organized S units tuned to the same preferred stimulus but at slightly different positions and scales and combined these subunits with a MAX-like operation, thereby introducing tolerance to scale and translation. The existence of a MAX operation in visual cortex was predicted by [5].
from theoretical arguments (and limited experimental evidence [31]) and was later supported experimentally in both V4 [32] and V1 at the complex cell level [33].

A gradual increase in both selectivity and scale (as observed along the ventral stream and as obtained in the model by interleaving the two key operations) is critical for avoiding both a combinatorial explosion in the number of units, and the binding problem between features. Below we shortly give idealized mathematical approximations of the operations.

Idealized mathematical descriptions of the two operations: In the following, we denote by $y$ the response of a unit (simple or complex). The set of inputs to the cell (i.e., presynaptic units) are denoted with subscripts $j = 1, \ldots, N$. When presented with a pattern of activity $x = (x_1, \ldots, x_N)$ as input, an idealized and static description of a complex unit response $y$ is given by:

$$y = \max_{j=1,\ldots,N} x_j$$

As mentioned above, for a complex cell, the inputs $x_j$ to the units are retinotopically organized (selected from an $m \times m$ grid of afferents with the same selectivity). For instance, in the case of a V1-like complex cell tuned to a horizontal bar, all input subunits are tuned to a horizontal bar but at slightly different positions and scales. Similarly, an idealized description of a simple unit response is given by:

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

$\sigma$ defines the sharpness of the Tuning of the unit around its preferred stimulus corresponding to the synaptic strengths $w = (w_1, \ldots, w_N)$. As for complex cells, the subunits of the simple cells are also retinotopically organized (selected from an $m \times m$ grid of possible afferents). In contrast with complex cells, the subunits of a simple cell have different selectivities to increase the complexity of the preferred stimulus. For instance, for $S$ units at the earliest layers, the subunits are V1-like complex cells at different preferred orientations. The response of a simple unit is maximal when the current pattern of input $x$ exactly matches the synaptic weights $w$ (for instance the frontal view of a face) and decreases with a bell-shaped profile as the pattern of input becomes more dissimilar (as the face is rotated away from the preferred view).

Both of these mathematical descriptions are only meant to describe the response behavior of cells at a phenomenological level. Plausible biophysical circuits for the Tuning and Max operations have been proposed based on feedforward and/or feedback shunting inhibition combined with normalization (see [34] and references therein).

2.2.2 Building a dictionary of shape-components from V1 to IT

The overall architecture is sketched in Figure 1 and reflects the general organization of visual cortex in a series of layers from V1 to IT and PFC. Colors encode the tentative correspondences between the functional primitives of the theory (right) and the structural primitives of the ventral stream in the primate visual system [19] (left, modified from [35]). Below we give a brief description of a model instantiating the theory. The reader should refer to [36] for a more complete description of the architecture and detailed parameter values.

The first stage of simple units ($S_1$) – corresponding to the classical simple cells of Hubel and Wiesel – represents the result of the first tuning operation: Each $S_1$ cell is tuned in a Gaussian-like way to a bar of a certain orientation among four possible ones. Each of the complex units
in the second layer \((C_1)\) – corresponding to the classical complex cells of Hubel & Wiesel – receives, within a neighborhood, the outputs of a group of simple units in the first layer at slightly different positions and sizes but with the same preferred orientation. The operation is a nonlinear max-like operation (see Eq. 1) which increases invariance to local changes in position and scale while maintaining feature specificity.

Please insert Figure 1 about here

At the next simple cell layer \((S_2)\), the units pool the activities of several complex units \((C_1)\) with weights dictated by the unsupervised learning stage (see below), yielding selectivity to more complex patterns such as combinations of oriented lines. Simple units in higher layers \((S_3 \text{ and } S_4)\) combine more and more complex features with a Gaussian tuning function (see Eq. 2), while the complex units \((C_2 \text{ and } C_3)\) pool their afferents through a max-like function (see Eq. 1) providing increasing invariance to position and scale. In the model, the two layers alternate (see [5]). Besides the main route that follows stages along the hierarchy of the ventral stream step-by-step, there exist several routes which bypass some of the stages, e.g., direct projections from V2 to IT (bypassing V4) and from V4 to IT (bypassing posterior IT cortex). In the model, such bypass route corresponds to the projections from the \(C_1\) layer to the \(S_{2b}\) and then \(C_{2b}\) layers. Altogether the various layers in the architecture – from \(V1\) to IT – create a large and redundant dictionary of features with different degrees of selectivity and invariance.

Although the present implementation follows the hierarchy of Figure 1, the hierarchy may not be as strict. For instance there may be units with relatively complex receptive fields already in \(V1\) [37, 38]. A mixture of cells with various levels of selectivity has also commonly been reported in \(V2, V4\) and IT [22]. In addition, it is likely that the same stimulus-driven learning mechanisms implemented for \(S_2\) units and above operate also at the level of the \(S_1\) units. This may generate \(S_1\) units with tuning not only for oriented bars but also for more complex patterns (e.g., corners), corresponding to the combination of LGN-like, center-surround subunits in specific geometrical arrangements. Indeed it may be advantageous for circuits in later stages (e.g., task-specific circuits in PFC) to have access not only to the highly invariant and selective units of AIT but also to less invariant and simpler units such as those in \(V2\) and \(V4\). Fine orientation discrimination tasks, for instance, may require information from lower levels of the hierarchy such as \(V1\). There might also be high level recognition tasks that benefit from less invariant representations.

2.2.3 Learning

Unsupervised developmental-like learning from \(V1\) to IT: Various lines of evidence suggest that visual experience – during and after development – together with genetic factors determine the connectivity and functional properties of cells in cortex. In this work, we assume that learning plays a key role in determining the wiring and the synaptic weights for the model units. We suggest that the tuning properties of simple units – at various levels in the hierarchy – correspond to learning which combinations of features appear most frequently in images. This is roughly equivalent to learning a dictionary of image patterns that appear with higher probability. The wiring of the \(S\) layers depends on learning correlations of features in the image which are present at the same time (i.e., for \(S_1\) units, the bar-like arrangements of LGN inputs, for \(S_2\) units, more complex arrangements of bar-like subunits, etc).

The wiring of complex cells, on the other hand, may reflect learning from visual experience to associate frequent transformations in time – such as translation and scale – of specific complex features coded by simple cells. The wiring of the \(C\) layers reflects learning correlations across time: e.g., at the \(C_1\) level, learning that afferent \(S_1\) units with the same orientation and
neighboring locations should be wired together because such a pattern often changes smoothly in time (under translation) [39]. Thus, learning at the $S$ and $C$ levels involves learning correlations present in the visual world. At present it is still unclear whether these two types of learning require different types of synaptic “rules” or not.

In the present model we only implemented learning at the $S$ level (connectivity at the $C$ level was hardwired based on physiology data). The goal of this learning stage is to determine the selectivity of the $S$ units, i.e., set the weight vector $\mathbf{w}$ (see Eq. 2) of the units in layers $S_2$ and higher. More precisely, the goal is to define the basic types of units in each of the $S$ layers which constitute a basic dictionary of shape-components with units that are tuned to image-features that occur with high probability in natural images. This assumption follows the notion that the visual system, through visual experience and evolution, may be adapted to the statistics of its natural environment [40]. Details about the learning rule can be found in [36].

**Supervised learning of the task-specific circuits from IT to PFC:** We assume that a particular program or routine is set up somewhere beyond IT (possibly in PFC [16, 41], but the exact locus may depend on the task). In a passive state (no specific visual task is set) there may be a default routine running (perhaps the routine: what is there?). Here we think of this routine as a particular PFC-like classification unit which combines the activity of a few hundred $S_4$ units tuned to examples of the target object (as well as distractors). While learning in the model from $S_2$ to $S_4$ is stimulus-driven, the PFC-like classification units are trained in a supervised way. The concept of a classifier that takes its inputs from a few broadly tuned example-based units is a learning scheme that is closely related to Radial Basis Function (RBF) networks [30], which are among the most powerful classifiers in terms of learning to generalize. Computer simulations have shown the plausibility of this scheme for visual recognition and its quantitative consistency with many data from physiology and psychophysics [42].

In the model, the response of a PFC-like classification unit with input weights $\mathbf{c} = (c_1, \ldots, c_n)$ is given by:

$$ f(\mathbf{x}) = \sum_i c_i K(\mathbf{x}^i, \mathbf{x}) \quad \text{where} \quad K(\mathbf{x}^i, \mathbf{x}) = \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^n (x^i_j - x_j)^2 \right). $$

$K(\mathbf{x}^i, \mathbf{x})$ characterizes the activity of the $i^{th}$ $S_4$ unit, tuned to the training example $\mathbf{x}^i$, in response to the input image $\mathbf{x}$ and was obtained by replacing the weight vector $\mathbf{w}$ in Eq. 2 by the training example $\mathbf{x}^i$ (i.e., $\mathbf{w} = \mathbf{x}^i$). The superscript $i$ indicates the index of the image in the training set and the subscript $j$ indicates the index of the pre-synaptic unit. Supervised learning at this stage involves adjusting the synaptic weights $\mathbf{c}$ to minimize the overall classification error on the training set (see [36]).

### 3 Comparison with physiological observations

The quantitative implementation of the model, as described in the previous section, allows for direct comparisons between the responses of units in the model and electrophysiological recordings from neurons in the visual cortex. Here we illustrate this approach by directly comparing the model against recordings from the macaque monkey inferior temporal cortex while the animal was passively viewing complex images.
3.1 Comparison of model units with physiological recordings in the ventral visual cortex

The model includes several layers that are meant to mimic visual areas V1, V2, V4 and IT cortex (Figure 1). We directly compared the responses of the model units against electrophysiological recordings obtained throughout all these visual areas. The model is able to account for many physiological observations in early visual areas. For instance, at the level of V1, model units agree with the tuning properties of cortical cells in terms of both frequency and orientation bandwidth, as well as peak frequency selectivity and receptive field sizes (see [43]). Also in V1, we observed that model units (corresponding to the $C_1$ layer) could explain responses of a subpopulation of complex cells obtained upon presenting two oriented bars within the receptive field [33]. At the level of V4, model $C_2$ units exhibit tuning for complex gratings (based on the recordings from [44]), and curvature (based on [45]), as well as interactions of multiple dots (based on [46]) or the simultaneous presentation of two-bar stimuli (based on [47], see [34] for details). For a more detailed comparison of the model against neuronal activity in multiple visual areas, see [34].

In sum, the model can capture many aspects of the physiological responses of neurons along the ventral visual stream from V1 to IT cortex. Thus, the input to the IT neurons, as used in the next section, constitutes a reasonable approximation of the physiological input to IT.

3.2 Decoding object information from IT and model units

We recently used a simple linear statistical classifier to quantitatively show that we could accurately, rapidly and robustly decode visual information about objects from the activity of small populations of neurons in anterior inferior temporal cortex [16]. In collaboration with Chou Hung and James DiCarlo at MIT, we observed that a binary response from the neurons (using small bins of 12.5 ms to count spikes) was sufficient to encode information with high accuracy. This robust visual information, as measured by our classifiers, could at least in principle be decoded by the targets of IT cortex such as prefrontal cortex [21]. Importantly, the population response generalized across object positions and scales. This scale and position invariance was evident even for novel objects that the animal never observed before (see also [27]). The observation that scale and position invariance occurs for novel objects strongly suggests that these two forms of invariance do not require multiple examples of each specific object. This should be contrasted with other forms of invariance, such as robustness to depth rotation, which requires multiple views in order to be able to generalize [30].

3.2.1 Read-out from $C_2b$ units is similar to decoding from IT neurons

We examined the responses of the model units to the same set of 77 complex object images seen by the monkey. These objects were divided into 8 possible categories. The model unit responses were divided into a training set and a test set. We used a one-versus-all approach, training 8 binary classifiers, one for each category against the rest of the categories, and then taking the classifier prediction to be the maximum among the 8 classifiers (for further details, see [16, 34]). Similar observations were made when trying to identify the objects by training 77 binary classifiers. For each layer of the model (see Figure 1), we randomly selected a subpopulation of model units for decoding. The input to the classifier consisted of the responses of the model units and the labels to the object categories (or object identities for the identification task). Data from multiple units were concatenated assuming independence.

We observed that we could accurately read out the object category and identity from model units. In Figure 2A, we compare the classification performance, for the categorization task
described above, between the IT neurons and the $C_{2b}$ model units. In agreement with the experimental data from IT, units from the $C_{2b}$ stage of the model yielded very high performance levels (performance $>70\%$ for 100 units). We observed that the physiological observations were in agreement with the predictions made by the highest layers in the model ($C_{2b}, S_4$) but not by earlier stages ($S_1$ through $S_2$). As expected, the layers from $S_1$ through $S_2$ showed a weaker degree of scale and position invariance.

The classification performance of $S_{2b}$ units (the input to $C_{2b}$ units, see Figure 1) was qualitatively close to the performance of local field potentials (LFPs) in IT cortex [48]. The main components of LFPs are dendritic potentials and therefore LFPs are generally considered to represent the dendritic input and local processing within a cortical area [49, 50]. Thus, it is tempting to speculate that the $S_{2b}$ responses in the model capture the type of information conveyed by LFPs in IT. However, care should be taken in this interpretation as the LFPs constitute an aggregate measure of the activity over multiple different types of neurons and large areas. Further investigation of the nature of the LFPs and their relation with the spiking responses could help unravel the transformations that take place across cortical layers.

The pattern of errors made by the classifier indicates that some groups were easier to discriminate than others. This was also evident upon examining the correlation matrix of the population responses for all pairs of pictures [16, 34]. The units yielded similar responses to stimuli that looked alike at the pixel level. The performance of the classifier for categorization dropped significantly upon arbitrarily defining the categories as random groups of pictures.

The performance values shown in Figure 2A are based on the responses of model units to single stimulus presentations that were not included in the classifier training and correspond to the results obtained using a linear classifier. Although the way in which the weights were learned (SVM) is probably very different in biology (see [36], once the weights are established the linear classification boundary could very easily be implemented by neuronal hardware (see Eq. 3). Therefore, the recognition performance provides a lower bound to what a real downstream unit (e.g., in PFC) could, in theory, perform on a single trial given input consisting of a few spikes from the neurons in IT cortex. Overall, we observed that the population of $C_{2b}$ model units yields a read-out performance level that is very similar to the one observed from a population of IT neurons.

3.2.2 Extrapolation to larger object sets

One of the remarkable aspects of primate visual recognition is the large number of different objects that can be identified. Although the exact limits are hard to bound, coarse estimates suggest that it is possible to visually recognize on the order of $10^4$ different concepts [51]. The physiological recordings were necessarily limited to a small set of objects due to time constraints during a recording session. Here we show that this type of encoding can extrapolate to (at least) reading out object category in a set consisting of 787 objects divided into 20 categories (the physiological observations and the model results discussed above were based on 77 objects divided into 8 categories).

The population of $C_{2b}$ units conveyed information that could be decoded to detect an object’s category across novel objects. The classifier was trained with objects from 20 possible categories presented at different random locations and the test set included novel objects never seen before by the classifier but from the same categories. These results show that a relatively small neuronal population can in principle support object recognition over large object sets. Similar results were obtained in analogous computer vision experiments using an even larger
set known as the Caltech101 object dataset [52, 53] where the model could perform object categorization between 101 categories. Other investigators have also used models that can extrapolate to large numbers of objects [54] or suggested that neuronal populations in IT cortex can also extrapolate to many objects [16, 55].

The number of objects (or classes) that can be decoded at a given level of accuracy grows approximately as an exponential function of the number of units. Even allowing for a strong redundancy in the number of units coding each type of feature, these results suggest that networks of thousands of units could display a very large capacity. Of course the argument above relies on several assumptions that could well be wrong. However, at the very least, these observations suggest that there does not seem to be any obvious capacity limitations for hierarchical models to encode realistically large numbers of objects and categories.

3.2.3 Robustness in object recognition

Many biological sources of noise could affect the encoding of information. Among the most drastic sources of noise are synaptic failures and neuronal death. To model this, we considered the performance of the classifier after randomly deleting a substantial fraction of the units during testing. As shown for the experimental data in [16], the classifier performance was very robust to this source of noise.

As discussed in the introduction, one of the main achievements of visual cortex is the balance of invariance and selectivity. Two particularly important forms of invariance are the robustness to changes in scale and position of the images. In order to analyze the degree of invariance to scale and position changes, we studied the responses of units at different stages of the model to scaled (0.5 × and 2 ×) and translated (2 degrees and 4 degrees) versions of the images. Figure 2A shows that the earlier stages of the model show a poor read-out performance under these transformations, but the performance of the \( C_{2b} \) stage is quite robust to these transformations as in the experimental data of Hung et al. in IT [16].

We also observed that the population response could extrapolate to novel objects within the same categories by training the classifier on the responses to 70% of the objects and testing its performance on the remaining 30% of the objects [34]. This suggests another dimension of robustness, namely, the possibility of learning about a category from some exemplars and then extrapolating for novel objects within the same category.

The results shown above correspond to randomly selecting a given number of units to train and test the classifier. The brain could be wired in a very specific manner so that only neurons highly specialized for a given task project to the neurons involved in decoding the information for that task. Pre-selecting the units (e.g., using those yielding the highest signal-to-noise ratio) yields similar results while using a significantly smaller number of units. Using a very specific set of neurons (instead of randomly pooling from the population and using more neurons for decoding) may show less robustness to neuronal death and spike failures. The bias towards using only a specific subset of neurons could be implemented through selection mechanisms including attention. For example, when searching for the car keys, the weights from some neurons could be adjusted so as to increase the signal-to-noise ratio for specific tasks. This may suggest that other concomitant recognition tasks would show weaker performance. In this case, the selection mechanisms take place before recognition by biasing specific populations for certain tasks.

3.2.4 Recognition in clutter

The decoding experiments described above, and a large fraction of the studies reported in the literature, involve the use of well-delimited single objects on a uniform background. This is
quite remote from natural vision where we typically encounter multiple objects embedded in different backgrounds, with potential occlusions, changes in illumination, etc.

Ultimately, we would like to be able to read out information from IT or from model units under natural vision scenarios in which an everyday life image can be presented and we can extract from the neurons/units population activity the same type and quality of information that a human observer can (in a flash). Here we show the degree of robustness of the decoding approach when objects are embedded in complex backgrounds (see also Section 4 describing the performance of the model in an animal vs. non-animal categorization task using objects embedded in complex backgrounds).

We presented the same 77 objects used in Figure 2A overlayed on top of images containing complex background scenes (Figure 2C). We did not attempt to make the resulting images realistic or meaningful in any way. Cognitive influences, memory and expectations play a role in object recognition. These high level effects may perhaps be mediated through feedback biasing mechanisms which would indicate that a monitor is more likely to be found on an office desk than in the jungle. However, the model described here is purely feedforward and does not include any of these potential biasing mechanisms. We used 4 different relative sizes of object to background (ratio of object area to whole image area) ranging from 6% to 69%. The latter condition is very similar to the single object situation analyzed above, both perceptually and in terms of the performance of the classifier. The smaller relative size makes it difficult to detect the object at least in some cases when it is not salient (see also Section 4). The classifier was trained on all objects using 20% of the background scenes and performance was evaluated using the same objects presented on the remaining novel background scenes (we used a total of 98 complex background scenes with photographs of outdoor scenes). The population of \( C_{2b} \) units allowed us to perform both object recognition (Figure 2C) and identification significantly above chance in spite of the background. Performance depended quite strongly on the relative image size (Figure 2C). The largest size (69%) yielded results that were indistinguishable from the single isolated object results discussed above (cf. Figure 2A). The small relative image size (6%) yielded comparatively lower results but the performance of \( C_{2b} \) units was still significantly above chance levels both for categorization as well as identification.

Recognizing (and searching for) small objects embedded in a large complex scene, (e.g., searching for the keys in your house), constitutes an example of a task that may require additional resources. These additional resources may involve serial attention which is likely to be dependent on feedback connections. Therefore, the model may suggest tasks and behaviors that require processes that are not predominantly feedforward.

### 3.2.5 Reading-out from images containing multiple objects

In order to further explore the mechanisms for representing information about an object’s identity and category in natural scenes, we studied the ability to read out information from the model units upon presentation of more than one object. We presented two objects simultaneously in each image (Figure 2B). During testing, the classifier was presented with images containing multiple objects. We asked two types of questions: (i) what is the most likely object in the image? and (ii) what are all the objects present in the image?

Training was initially performed with single objects. Interestingly, we could also train the classifier using images containing multiple objects. In this case, for each image, the label was the identity (or category) of one of the objects (randomly chosen so that the overall training set had the same number of examples for each of the objects or object categories). This is arguably a more natural situation under which we learn about objects since we rarely see isolated objects. It is possible that attentional biases to some extent “isolate” an object (e.g., when learning about
an object with an instructor that points to it).

We first considered the best prediction of the classifier (as we did above for Figure 2A). We defined a hit in the output of the classifier if the classifier correctly predicted either one of the two objects presented during testing. The population of $C_{2b}$ model units yielded very high performance reaching more than 90% both for categorization and identification with the single object training and reaching more than 80% with the multiple object training. Given that in each trial there are basically two possibilities to get a hit, the chance levels are higher than the ones reported in Figure 2A. However, it is clear that the performance of the $C_{2b}$ population response is significantly above chance indicating that accurate object information can be read-out even in the presence of another object. We also extended these observations to 3 objects and to 10 objects [34], obtaining qualitatively similar conclusions.

Ultimately, we would like to be able to understand an image in its entirety, including a description of all its objects. Therefore, we asked a more difficult question by requiring the classifier to correctly predict all the objects (or all the object categories) present in the image. We generally live under the illusion that we can recognize and describe every object in an image during a glimpse. Multiple psychophysics studies suggest that this is probably wrong. Perhaps one of the most striking demonstrations of this fallacy is the fact that sometimes we can be oblivious to large changes in the images (see for example [56]). What is the capacity of the representation at-a-glance? There is no consensus answer to this question but some psychophysical studies suggest that only a handful of objects can be described in a brief glimpse of an image (on the order of 5 objects). After this first glance, eye movements and/or attentional shifts may be required to further describe an image. We continue here referring to this rapid vision scenario and we strive to explain our perceptual capabilities during the glance using the model. Thus, the goal is to be able to fully describe a set of about 5 objects that can be simultaneously presented in multiple backgrounds in a natural scenario.

For this purpose, we took the two most likely objects (or object categories) given by the two best classifier predictions. A hit from the classifier output was defined as a perfect match between these predictions and the two objects present in the image. This task is much more difficult (compared to the task where the goal is to categorize or identify any of the objects in the image). The performance of the classifier was also much smaller than the one reported for the single-object predictions. However, performance was significantly above chance, reaching almost 40% for categorization (chance $= 0.0357$) and almost 8% for identification (chance $= 3.4 \times 10^{-4}$).

Similar results were obtained upon reading out the category or identity of all objects present in the image in the case of 3-object and 10-object images. Briefly, even in images containing 10 objects, it is possible to reliably identify one arbitrary object significantly above chance from the model units. However, the model performance in trying to describe all objects in the image drops drastically with multiple objects to very low levels for 4-5 objects.

In summary, these observations suggest that it is possible to recognize objects from the activity of small populations of IT-like model units under natural situations involving complex backgrounds and several objects. The observations also suggest that, in order to fully describe an image containing many objects, eye movements, feedback or other additional mechanisms may be required.

4 Performance on Natural Images

Linking physiology to behavior is a critical step towards understanding visual cortex. The architecture described in Section 2 (Figure 1) attempts to account for the known properties of neurons along the ventral stream of visual cortex as closely as possible. For a theory of visual
cortex to be successful, it should not only be able to mimic the response properties of neurons to artificial stimuli like the ones typically used in a physiology or a psychology lab but also to be able to perform complex categorization tasks in a real-world setting.

4.1 Comparison between the model and computer vision systems

We extensively tested the model on standard computer vision databases for comparison with several state-of-the-art AI systems (see [36, 52, 53] for details). Such real-world image datasets tend to be much more challenging than the typical ones used in the neuroscience lab; They usually involve different object categories and the systems that are evaluated have to cope with large variations in shape, contrast, clutter, pose, illumination, size, etc.). Given the many specific biological constraints that the theory had to satisfy (e.g., using only biophysically plausible operations, receptive field sizes, range of invariances, etc.) it was not clear how well the model implementation described in Section 2 would perform in comparison to systems that have been heuristically engineered for the task.

Surprisingly we found that the model is capable of recognizing complex images (see [52, 53]). For instance, the model performs at a level comparable to some of the best existing systems on the CalTech-101 image database of 101 object categories [57] with a recognition rate of almost 60% (chance level < 1%), see [58, 59]. Additionally, Bileschi & Wolfe developed an automated real-world Street Scene recognition system [60] based on the model described in Section 2. The system is able to recognize well seven different object categories (cars, bikes, cars, skies, roads, buildings, trees) from natural images of street scenes despite very large variations in shape (e.g., trees in summer and winter, SUVs as well as compact cars under any view point).

Because of the good agreement between the tuning properties of units in various stages of the model and neurons in corresponding brain areas (see Section 3) and because of the ability of the model to perform complex recognition tasks with natural images, we tested whether the level of performance achieved by the model was sufficient to account for the level of performance of human observers. To test this hypothesis, in the same way as an experimental test of Newton’s second law requires choosing a situation in which friction is negligible, we looked for an experimental paradigm in which recognition has to be fast and cortical back-projections are likely to be inactive. Ultra-rapid object categorization [13] likely depends only on feedforward processing [13, 15, 17, 61, 62] and thus satisfies our criterion.

4.2 Comparison between the model and human observers

Human observers can discriminate a scene that contains a particular prominent object, like an animal or a vehicle after only 20 ms of exposure; evoked response potential components related to either low-level features of the image categories (e.g., animal or vehicles) or to the image status (animal present or absent) are available at 80 and 150 ms respectively. These experimental results establish an upper bound on how fast visual categorization decisions can be made by the human visual system, and suggest that categorical decisions can be implemented within a feedforward mechanism of information processing [13, 15, 17, 61, 62].

4.2.1 The model accounts for the level of performance of human observers on a rapid animal vs. non-animal categorization task

**Methods:** In collaboration with Aude Oliva at MIT, we tested human observers on a rapid animal vs. non-animal categorization task (see [63] for details). The choice of the animal category was motivated by the fact that 1) it was used in the original paradigm by Thorpe and
colleagues [13] and 2) animal stimuli constitute a rich class of stimuli exhibiting large variations in texture, shape, size, etc such that no computer vision systems have been tested on thus far.

We used an image dataset that was collected by Antonio Torralba and Aude Oliva and consisted in a balanced set of 600 animal and 600 non-animal images (see [64]). The 600 animal images were selected from a commercially available database (Corel Photodisc) and grouped into four categories, each category corresponding to a different viewing-distance from the camera: heads (close-ups), close-body (animal body occupying the whole image), medium-body (animal in scene context) and far-body (small animal or groups of animals in larger context). One example from each group is shown in Figure 3.

To make the task harder and prevent subjects from relying on low-level cues such as image-depth, the six hundred distractor images were carefully selected to match each of the four viewing-distances. Distractor images were of two types (three hundred of each): artificial or natural scenes. Images were all converted to gray values. Figure 3 shows typical examples of the stimuli used in this experiment (see [63] for details).

During the experiment, images were briefly flashed for 20 ms, followed by an inter-stimulus interval (i.e., a blank screen) of 30 ms, followed by a mask (80 ms, 1/f noise). This is usually considered a long stimulus onset asynchrony (SOA = 50 ms) for which human observers are close to ceiling performance [65]. On the other hand, based on latencies in visual cortex (see [63]) such SOA should minimize the possibility of feedback and top-down effects in the task: We estimated from physiology data (see [63]) that feedbacks from say, V4 to V1 or IT/PFC to V4, should occur not earlier than 40 – 60 ms after stimulus onset. Human observers (nh = 24) were asked to respond as fast as they could to the presence or absence of an animal in the image by pressing either of two keys.

Before we could evaluate the performance of the model, the task-specific circuits from IT to PFC (see Section 2) had to be trained. These task-specific circuits correspond to a simple linear classifier that reads out the activity of a population of high-level model units analogous to recordings from anterior IT cortex (see Section 3). The training for these task-specific circuits was done by using (nm = 20) random splits of the 1,200 stimuli into a training set of 600 images and a test set of 600 images. For each split, we learned the synaptic weights of the task-specific circuits of the model by minimizing the error on the training set (see [63]) and evaluated the model performance on the test set. The reported performance corresponds to the average performance from the (nm = 20) random runs.

Results: The performance of the model and of human observers was very similar (see Figure 3). As for the model, human observers performed best on “close-body” views and worst on “far-body” views. An intermediate level of performance was obtained for “head” and “medium-far” views. Overall no significant difference was found between the level of performance of the model and human subjects. Interestingly, the observed dependency between the level of performance and the amount of clutter in the images (which increases from the close-body to the far-body condition) for both human observers and the model seems consistent with the read-out experiment from IT neurons (for both the model and human observers) as described in Section 3.

Importantly, lower stages of the model (C1 units) alone could not account for the results (see [63]). Additionally lesioning V4 in the model, (i.e., leaving the bypass routes (C2b units as the only source of inputs to the final classifier, see Figure 1) also resulted in a significant loss in performance (this was true even after retraining the task-specific circuits thus accounting for a “recovery” period). This lesion experiment suggests that the large dictionary of shape-tuned units in the model (from V1 to IT) with different levels of complexity and invariance learned
from natural images is key in explaining the level of performance.

Beyond comparing levels of performance, we also performed an image-by-image comparison between the model and human observers. For this comparison, we defined an index of “animalness” for each individual images. For the model, this index was computed by calculating the percentage of times each image was classified as an animal (irrespective of its true label) for each random run ($n_m = 20$) during which it was presented as a test image. For human observers we computed the number of times each individual image was classified as an animal by each observer ($n_h = 24$). This index measures the confidence of either the model ($n_m = 20$) or human observers ($n_h = 24$) in the presence of an animal in the image. A percentage of 100% (correspondingly 0%) indicates a very high level of confidence in the presence (absence) of an animal. The level of correlation for the animalness index between the model and human observers was 0.71, 0.84, 0.71 and 0.60 for heads, close-body, medium-body and far-body respectively ($p < 0.01$ for testing the hypothesis of no correlation against the alternative that there is a non-zero correlation). This suggests that the model and human observers tend to produce consistent responses on individual images.

4.2.2 Comparison between the model and human observers on rotated images

Additionally, to further challenge the model, we looked at the effect of image orientation ($90^\circ$ and $180^\circ$ in-the-plane rotation): Rousselet et al. previously suggested that the level of performance of human observers during a rapid categorization task tend to be robust to image rotation [66]. We wanted to test this hypothesis on the model and found that the model and human observers exhibited a similar degree of robustness (see [63]). Interestingly, the good performance of the model on rotated images was obtained without the need for retraining the model. This suggests that according to the dictionary of shape-tuned units from V1 to IT in the model (and presumably in visual cortex), an image of a rotated animal is more similar to an image of an upright animal than to distractors. In other word a small image patch of a rotated animal is more similar to a patch of an upright animal than to a patch of image from a distractor.

4.2.3 Discussion: Feedforward vs. feedback processing

As discussed earlier, an important assumption for the experiment described above is that with an SOA of 50 ms, the mask leaves sufficient time to process the signal and estimate firing rates at each stage of the hierarchy (i.e., 20-50 ms, see [15, 61, 67, 68]) yet selectively blocks top-down signals (e.g., from IT or PFC to V4 that we estimated to be around 40-60 ms), see [63] for a complete discussion. The prediction is thus that the feedforward system should: 1) outperform human observers for very short SOAs (i.e., under 50 ms when there is not enough time to reliably perform local computations or estimate firing rates within visual areas), 2) mimic the level of performance of human observers for SOAs around 50 ms such that there is enough time to reliably estimate firing rates within visual areas but not enough time for back-projections from top-down to become active and 3) underperform human observers for long SOAs (beyond 60 ms) such that feedbacks are active.

We thus tested the influence of the mask on visual processing with four experimental conditions, i.e., when the mask followed the target image a) without any delay (with an SOA of 20 ms), b) with an SOA of 50 ms (corresponding to an inter-stimulus interval of 30 ms), c) with an SOA of 80 ms or d) never (“no-mask” condition). For all four conditions, the target presentation was fixed to 20 ms as before. As expected, the delay between the stimulus and the
mask onset modulates the level of performance of the observers, improving gradually from the 20 ms SOA condition to the no-mask condition. The performance of the model was superior to the performance of human observers for the SOA of 20 ms. The model closely mimicked the level of performance of human observers for the 50 ms condition (see Figure 3). The implication would be that, under these conditions, the present feedforward version of the model already provides a satisfactory description of information processing in the ventral stream of visual cortex. Human observers however outperformed the model for the 80 ms SOA and the no-mask condition.

5 Discussion

5.1 General remarks about the theory

We have developed a quantitative model of the feedforward pathway of the ventral stream in visual cortex – from cortical area V1 to V2 to V4 to IT and PFC – that captures its ability to learn visual tasks, such as identification and categorization of objects from images. The quantitative nature of the model has allowed us to directly compare its performance against experimental observations at different scales and also against current computer vision algorithms. Here we showed how the model can explain experimental results from visual object recognition within short times at two very different scales: from human psychophysics to physiological recordings in IT cortex. The model certainly does not account for all possible visual phenomena and illusions (see also extensions, predictions and future directions below). However, the success of the model in explaining experimental data across multiple scales and making quantitative predictions strongly suggests that the theory provides an important framework for the investigation of the feedforward path in visual cortex and the processes involved in immediate recognition.

An important component of a theory is that it should be possible to falsify it. In that spirit, we list some key experiments and findings here that could refute the present framework. First, a strong dissociation between experimental observations and model predictions would suggest that revisions need to be made to the model (e.g., psychophysical or physiological observations that cannot be explained or contradict predictions made by the model). Second, as stated in the introduction, the present framework relies entirely on a feedforward architecture from V1 to IT and PFC. Any evidence that feedback plays a key role during the early stages of recognition should be considered as hard evidence suggesting that important revisions would need to be made in the main architecture of the model (Figure 1).

5.2 A wish-list of experiments

Here we discuss some predictions from the theory and an accompanying “wish list” of experiments that could be done to test, refute or validate those predictions. We try to focus on what we naively think are relatively feasible experiments.

1. The distinction between simple and complex cells has been made only in primary visual cortex. Our theory and parsimony considerations suggest that a similar circuit is repeated throughout visual cortex. Therefore, unbiased recordings from neurons in higher visual areas may reveal the existence of two classes of neurons which could be distinguished by their degree of invariance to image transformations.

2. As the examples discussed in this manuscript illustrate, our theory can make quantitative predictions about the limits of immediate recognition at the behavioral level (Section 4) and also at the neuronal level (Section 3). Some of the conditions that pose the biggest
challenges include recognition of objects that are small relative to the whole image and also the presence of multiple objects, background and clutter. It would be of interest to compare these predictions against behavioral and physiological measurements. This could be achieved by adding extra conditions in the psychophysical experiment of Section 4 and by extending the read-out experiments from Section 3 to natural images and more complex recognition scenarios.

3. The theory suggests that immediate recognition may rely on a large dictionary of shape-components (i.e., common image-features) with different levels of complexity and invariance. This fits well with the concept of “unbound features”[69, 70] postulated by cognitive theories of pre-attentive vision. Importantly the theory does not rely on any figure-ground separation. This suggests that, at least for immediate recognition, recognition can work without an intermediate segmentation step. Furthermore, it also suggests that it is not necessary to define objects as fundamental units in visual recognition.

4. There is no specific computational role for a functional topography of units in the model. Thus, the strong degree of topography present throughout cortex, may arise from developmental reasons and physical constraints (a given axon may be more likely to target two adjacent neurons than two neurons that are far away; also, there may be a strong pressure to minimize wiring) as opposed to having a specific role in object recognition or the computations made in cortex.

5. The response of a given simple unit in the model can be described by Eq. 2. Thus, there are multiple different inputs that could activate a particular unit. This may explain the somewhat puzzling observations of why physiologists often find neurons that seem to respond to apparently dissimilar objects. Following this reasoning, it should be possible to generate an iso-response stimulus set, i.e., a series of stimuli that should elicit similar responses in a given unit even when the stimuli apparently look different or the shape of the iso-response stimulus set appear non-intuitive.

6. It is tempting to antropomorphically attempt to assign specific words in English to the type of images that activate a particular unit. This has been carried to the extreme of even speaking of a neuron’s “preferences”. The current theory suggests that the input that can give rise to high responses from a neuron is at the same time simpler and more complex than this. It is simpler because it can be rigorously approximated by specific simple equations that control its output. It is more complex because these weight vectors and equations may not easily map to words such as “face neuron”, “curvature”, etc. Along these lines, the previous point also emphasizes that dissimilar stimuli can give rise to similar responses.

7. There are many tasks that may not require back-projections. The model should fail to perform attention-demanding tasks. Thus, the model’s performance could provide a reliable signature of whether a task can be accomplished during immediate recognition, in the absence of feedback (e.g., [17]). As stated above, one of the main assumptions of the current model is the feedforward architecture. This suggests that the model may not perform well in situations that require multiple fixations, eye movements and feedback mechanisms. Recent psychophysical work suggests that performance on dual tasks can provide a diagnostic tool for characterizing tasks that do or do not involve attention [17]. Can the model perform these dual tasks when psychophysics suggests that attention is or is not required? Are back-projections and feedback required?
In addition to the predictions listed above, we recently discussed other experiments and predictions which are based on a more detailed discussion of the biophysical circuits implementing the main operations in the model (see [34]).

5.3 Future directions

We end this article by reflecting on several of the open questions, unexplained phenomena and missing components of the theory. Before we begin, we should note that visual recognition encompasses much more than what has been attempted and achieved with the current theory. A simple example may illustrate this point. In the animal categorization task discussed in the previous section, humans make mistakes upon being pressed to respond promptly. Given 10 seconds and no mask, performance would be basically 100%. As stated several times, the goal here is to provide a framework to quantitatively think about the initial steps in vision but it is clear that much remains to be understood beyond immediate recognition.

5.3.1 Open questions

**How strict is the hierarchy and how precisely does it map into cells of different visual areas?**

For instance, are cells corresponding to \( S_2 \) units in V2 and \( C_2 \) units in V4 or are some cells corresponding to \( S_2 \) units already in V1? The theory is rather open about these possibilities: the mapping of Figure 1 is just an educated guess. However, because of the increasing arborization of cells and the number of boutons from V1 to PFC [71], the number of subunits to the cells should increase and thus their potential size and complexity. In addition, \( C \) units should show more invariance from the bottom to the top of the hierarchy.

**What is the nature of the cortical and subcortical connections (both feedforward and feedback) of the main areas of the ventral visual stream that are involved in the model?**

A more thorough characterization at the anatomical level of the circuits in visual cortex would lead to a more realistic architecture of the model by better constraining some of the parameters such as the size of the dictionary of shape-components or the number of inputs to units in different layers. This would also help refine and extend the existing literature on the organization of visual cortex [19]. With the recent development of higher resolution tracers (e.g., PHA-L, biocytin, DBA), visualization has greatly improved and it is now possible to go beyond a general layout of interconnected structures and start addressing the finer organization of connections.

**What are the precise biophysical mechanisms for the learning rule described in Section 2 and how can invariances be learned within the same framework?**

Possible synaptic mechanisms for learning should be described in biophysical details. As suggested earlier, there should be at least three different synaptic rules: 1) for learning the tuning of the units at the \( S \) level by detecting correlations among subunits at the same time; 2) for learning the invariance to position and scale at the \( C \) level by detecting correlations among subunits across time and 3) for training the task-specific circuits (probably from IT to PFC) in a supervised fashion.

**Is learning in areas below IT purely unsupervised and developmental-like as assumed in the theory?**

Or is there task- and/or object-specific learning in adults occurring below IT in V4, V2 or even V1.
Have we reached the limit of what feedforward architectures can achieve in terms of performance? In other words, is the somewhat better performance of humans on the animal vs. non-animal categorization task over the model for SOAs longer than 80 ms due to feedback effects mediated by the back-projections or is it that the model still need to be improved to attain human performance in the absence of a mask? There could be several directions to follow in order to try to improve the model performance. One possibility would involve experimenting with the size of the dictionary of shape-components (that could be further reduced with feature selection techniques for instance). Another possibility would involve adding intermediate layers to the existing ones.

Are feedback loops always desirable? Is the performance on a specific task guaranteed to always increase when subjects are given more time? Or are there tasks for which blocking the effect of back-projections with rapid masked visual presentation increases the level of performance compared to longer presentation times?

5.3.2 Future extensions

Learning the tuning of the $S_1$ units. In the present implementation of the model the tuning of the simple cells in V1 is hardwired. It is likely that it could be determined through the same passive learning mechanisms postulated for the $S_2$, $S_{2b}$, and $S_3$ units (possibly in V4 and PIT), possibly with a slower time scale and constrained to LGN center-surround subunits. We would expect the automatic learning from natural images mostly of oriented receptive fields but also of more complex ones, including end-stopping units (as reported for instance in [72] in layer 6 of V1).

Dynamics of neuronal responses The current implementation is completely static. This is very far from the brain and also precludes us from asking several questions about the encoding of the information, about learning, about the effect of time, about relative timing across areas, etc. Perhaps the easiest way to solve this is by using simple single neuron models (such as an integrate-and-fire neuron) for the units in the model. This question is clearly related to the biophysics of the circuitry, i.e., what type of biological architectures and mechanisms can give rise to the global operations used by the model. A dynamical model would allow us to more realistically compare to experimental data. For example, the experiments described in Section 4 compare the results in a categorization task between the model and human subjects. In the human psychophysics, the stimuli were masked briefly after stimulus presentation. A dynamical model would allow us to investigate the role and mechanisms responsible for masking. A dynamical model may also allow investigation of time-dependent phenomena as well as learning based on correlations across time.

Extensions of the model to other visual inputs There are many aspects of vision that are not currently implemented in the model. These include color, stereo, motion and time-varying stimuli. Initial work has been done to extend the model to the visual recognition of action and motions [73, 74]. It is likely that the same units supporting recognition of static images (the $S_4$, view-tuned units in the model) show time sequence selectivity. Color mechanisms from V1 to IT should be included. The present implementation deals with gray level images. This probably includes a simpler part which accounts for the fact that we can recognize the same object under different colors. More complex phenomena involving color such as color constancy and the influence of the background and integration in color perception should ultimately also be explained.
Stereo mechanisms from V1 to IT should also be included. Stereo and especially motion play an important role in the learning of invariances such as position and size invariance via a correlation-based rule such as the trace rule.

Extensions of the anatomy of the model  Even staying within the feedforward skeleton outlined here, there are many connections that are known to exist in the brain which are not accounted for in the current model. The goal of the model is to extract the basic principles used in recognition and not to copy, neuron by neuron, the entire brain. However, certain connectivity patterns may have important computational consequences. For example, there are horizontal connections in cortex that may be important in modulating and integrating information across areas beyond the receptive field.

Beyond a feedforward model  It has been known for many decades now that there are abundant back-projections in the brain. In the visual system, every area projects back to its input area (with the exception of the lateral geniculate nucleus in the thalamus which does not project back to the retina). Some of these connections (e.g., from V2 to V1), may play a role even during immediate recognition. However, a central assumption of the current model is that long-range backprojections (e.g., from area IT to V1) do not play a role during the first 100 to 150 ms of vision. Given enough time, humans make eye movements to scan an image and performance in many object recognition tasks can increase significantly over that obtained during fast presentation.

Visual illusions  A variety of visual illusions show striking effects that are often counterintuitive and require an explanation in terms of the neuronal circuits. While in some cases specific models have been proposed to explain one or another phenomena, it would be interesting to explore how well the model (and thus feedforward vision) can account for those observations. A few simple examples include illusory contours (such as the Kanizsa triangle), long-range integration effects (such as the Cornsweet illusion), etc. More generally, it is likely that early Gestalt-like mechanisms – for detecting collinearity, symmetry, parallelism, etc – exist in V1 or V2 or V4. They are not present in this version of the model. It is an open and interesting question how they could be added to it in a plausible way.

Acknowledgements

We would like to acknowledge Chou Hung and James DiCarlo for their contributions in the physiology experiments described in Section 3. We would like to acknowledge Aude Oliva for her contribution to the psychophysics experiments described in Section 4. We thank Timothee Masquellier for useful comments on this manuscript. This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, McGovern fellowship (GK), National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032

References


6 Figure Captions

Figure 1. Tentative mapping between (right) functional primitives of the theory and (left) structural primitives of the ventral stream in the primate visual system [19]. Colors encode the tentative correspondences between model layers and brain areas. Stages of simple cells with Gaussian-like tuning (plain circles and arrows), which provide generalization [42], are interleaved with layers of complex units (dotted circles and arrows), which perform a softmax operation on their inputs and provide invariance to position and scale (pooling over scales is not shown in the figure). Both operations may be performed by the same local recurrent circuits of lateral inhibition (see text). It is important to point out that the hierarchy is probably not as strict as depicted here. In addition there may be cells with relatively complex receptive fields already in V1. The main route from the feedforward ventral pathway is denoted with black arrows while the bypass route [75] is denoted with yellow arrows. Learning in the simple unit layers from V2/V4 up to IT (including the S4 view-tuned units) is assumed to be stimulus-driven. It only depends on task-independent visual experience tuning of the units. Supervised learning occurs at the level of the circuits in PFC (two sets of possible circuits for two of the many different recognition tasks – identification and categorization – are indicated in the figure at the level of PFC). The model, which is feedforward (apart from local recurrent circuits), attempts to describe the initial stage of visual processing, immediate recognition, corresponding to the output of the top of the hierarchy and to the first 150 milliseconds in visual recognition.

Figure 2. A. Classification performance for spiking activity from IT neurons (black) and C2b units from the model (gray). The performance shown here is based on the categorization task where the classifier was trained based on the category of the object. A linear classifier classifier was trained using the responses to the 77 objects at a single scale and position (shown for one objects by “TRAIN”). The classifier performance was evaluated using shifted or scaled versions of the same 77 objects (shown for one object by “TEST”). During training, the classifier was never presented with the unit responses to the shifted or scaled objects. The left-most column shows the performance for training and testing on separate repetitions of the objects at the same standard position and scale (this is shown only for the IT neurons because there is no variability in the model which is deterministic). The second bar shows the performance after training on the standard position and scale (3.4 degrees, center of gaze) and testing on the shifted and scaled images. The dashed horizontal line indicates chance performance (12.5 %, 1 out of 8 possible categories). Error bars show SD for 20 random choices of the units used for training/testing. B. Classification performance for reading out object category in the presence of two objects. We exhaustively studied all possible pairs using the same 77 objects as in part A (see two examples on the upper left part of the figure). The classifier was trained with images containing two objects and the label corresponded to the category of one of them. During testing, the classifier’s prediction was considered to be a hit if it correctly categorized either of the objects present in the image. The dashed line indicates chance performance obtained by randomly assigning object labels during training. C. Classification performance for reading out object category as a function of the relative size (area ratio) of object to background. Here the classifier was trained using the responses of 256 units to the objects presented in cluttered backgrounds. The classifier performance was evaluated using the same objects embedded in different backgrounds. The horizontal dashed line indicates chance performance obtained by randomly shuffling the object labels during training.

Figure 3. Comparison between the model and human observers. Images showed either an animal embedded in a natural background or a natural scene without any animals. Images
were flashed for 20 ms followed by a 30 ms blank and a 80 ms mask. Human observers or the model were queried to respond indicating whether an animal was present or not. The figure shows the accuracy as $d'$ (the higher the value of the $d'$, the better the performance), for the model (red) and humans (blue) across 1,200 animal and non-animal stimuli.
FIGURE 1
Figure 2

A.

Classification performance

Size: 3.4°  3.4°  1.7°  6.8°  3.4°  3.4°
Position: center  center  center  2° horz.  4° horz.

TRAIN

TEST

B.

Classification performance vs. Number of units

C.

Classification Performance vs. Area Ratio
FIGURE 3

Performance (d')

- Model
- Human-observers

Head | Close-body | Medium-body | Far-body

Images of animal heads and bodies.