SI Text

Supplementary web material can be found online [1] and includes, in particular, a basic software implementation and the stimulus database. It also includes a summary of various performance measures for both human observers and the model (including ROC analysis, error and hit rates) as well as the reaction times for human observers.

Model Overview

Layers in the model are organized in feature maps which may be thought of as columns or clusters of units with the same selectivity (or preferred stimulus) but with receptive fields at slightly different scales and positions (see SI Fig. 8). Within one feature map all units share the same selectivity, that is, the same synaptic weight vector $\mathbf{w}$ which is learned from natural images (see *Unsupervised Learning from V2 to IT*). As in the original Hubel & Wiesel proposal, there are two kinds of units / layers in the model:

- The S layers are composed of simple units. The pooling operation is a Gaussian-like tuning operation (see ref. [2] for an overview). That is, the response y of a simple unit, receiving the pattern of synaptic inputs $\left(x_1, \ldots, x_{n_{S_k}}\right)$ from the previous layer is given by:

$$y = \exp -\frac{1}{2\sigma^2}\sum_{j=1}^{n_{S_k}}(w_j - x_j)^2, \qquad (1)$$

  where σ defines the sharpness of the tuning around the preferred stimulus of the unit corresponding to the weight vector $\mathbf{w} = (w_1, \ldots w_{n_{S_k}})$.[1]

- The C layers are composed of complex units. The pooling operation is a max operation. That is, the response y of a simple unit, receiving the pattern of synaptic inputs $\left(x_1, \ldots, x_{n_{C_k}}\right)$ from the previous layer is given by:

$$y = \max_{j=1\ldots\ n_{C_k}} x_j. \qquad (2)$$

In this paper we used the static idealized operations described in Eq. 2 and Eq. 1. There are plausible local circuits [4] implementing the two key operations within the time constraints of the experimental data [5, 6] based on small local population of spiking neurons firing probabilistically in proportion to the underlying analog value [7] and on shunting inhibition [8]. Other possibilities may involve spike timing in individual neurons (see ref. [9] for a recent review). A complete description of the two operations, a summary of the evidence as well as plausible biophysical circuits to implement them can be found [3] (see Section 5, pp. 53-59).

There are several parameters governing the organization of individual layers: K_X is the number of feature maps in layer X. Units in layer X receive their inputs from a topologically related $\Delta N_X \times \Delta N_X \times \Delta S_X$ grid of possible afferent units from the previous layer where ΔN_X defines a range of positions and ΔS_X a range of scales (see SI Fig. 8 and SI Table 1 for the parameter values used in the current implementation). Simple units pool over afferent units at the same scale, that is, ΔS_{S_k} contains only a single scale element. Also, in the current model implementation, while complex units pool over all possible afferents such that each unit in layer C_k receives $n_{C_k} = \Delta N^S_{C_k} \times \Delta N^S_{C_k} \times \Delta S_{C_k}$, simple units receive only a subset of the possible afferent units (selected at random) such that $n_{S_k} < \Delta N_{S_k} \times \Delta N_{S_k}$ (see SI Table 1 for parameter values).

There exists a high degree of overlap between units in all stages. The number of feature maps is conserved from S_k to C_k stages, that is, $K_{S_k} = K_{C_k}$. There is a downsampling stage from S_k to C_k stage. While S units are computed at all possible locations, C units are only computed every ϵ_{C_k} possible locations.

S_1 units

The input to the model is a gray-value image ($256 \times 256 \sim 7^o \times 7^o$ of visual angle) which is first analyzed by a multi-dimensional array of simple S_1 units (see SI Fig. 9). S_1 units correspond to the simple cells of Hubel & Wiesel from striate cortex. The population of S_1 units consists in 96 types of units, that is, 2 phases $\times 4$ orientations $\times 17$ sizes (or equivalently peak spatial frequencies). The receptive field sizes of the S_1 units are in the same range as in primate visual cortex (that is, $0.2^o - 1.0^o$, see ref. [10, 11]). Peak frequencies are in the range $1.6 - 9.8$ cycles/deg. Details about the implementation of the S_1 units and their comparison with primate cortical cells can be found in [12].

Mathematically the weight vector $\mathbf{w}$ of the S_1 units take the form of a Gabor function which have been shown to provide a good model of simple cell receptive fields [13] and can be described by the following equation:

$$\begin{aligned} F(u_1, u_2) &= \exp\left(-\frac{(\hat{u_1}^2 + \gamma^2\hat{u_2}^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\hat{u_1}\right) \qquad (3) \\ \text{s.t.} & \\ \hat{u_1} &= u_1\cos\theta + u_2\sin\theta \\ \hat{u_2} &= -u_1\sin\theta + u_2\cos\theta. \end{aligned}$$

The five parameters, that is, 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 S_1 units. In setting these parameters we tried to generate a population of units that match the bulk of parafoveal cells as closely as possible (see ref. [12]).

C_1 units

The next C_1 level corresponds to striate complex cells [14]. Each of the complex C_1 units receives the outputs of a group of simple S_1 units with the same preferred orientation (and two opposite phases) but at slightly different positions and sizes (or peak frequencies). The result of the pooling over positions is that C_1 units become insensitive to the location of the stimulus within their receptive fields, which is a hallmark of the complex cells [14]. As a result the size of the receptive fields increase from the S_1 to the C_1 stage (from $0.2^o - 1.0^o$ to $0.4^o - 2.0^o$). Similarly the effect of the pooling over scales is a broadening of the frequency bandwidth from S_1 to C_1 units also in agreement with physiology [14, 10, 15]. The parameters of the C_1 units (see SI Table 1) were adjusted so as to match as closely as possible the tuning properties of V1 parafoveal complex cells [16].

S_2 units

At the next S_2 level, units pool the activities of $n_{S_2} = 10$ retinotopically organized complex C_1 units at different preferred orientations

[1] When Eq. 1 is approximated by a normalized dot-product followed by a sigmoid, i.e., $y = \frac{\sum_{j=1}^{n_{S_k}} w_j\, x_j^p}{k + (\sum_{j=1}^{n} x_j^q)^r}$, the weight vector $\mathbf{w}$ corresponds to the strength of the synaptic inputs to the Gaussian-tuned unit (see ref. [3], pp. 11-13).

over a $\Delta N_{S_2} \times \Delta N_{S_2} = 3 \times 3$ neighborhood of C_1 units via a tuning operation (see Eq.1). As a result, the complexity of the preferred stimuli is increased: At the C_1 level units are selective for single bars at a particular orientation, whereas at the S_2 level, units become selective to more complex patterns – such as the combination of oriented bars to form contours or boundary-conformations. Receptive field sizes at the S_2 level range between $0.6^o - 2.4^o$.

C_2 units

In the next C_2 stage units pool over S_2 units that are tuned to the same preferred stimulus (they correspond to the same combination of C_1 units and therefore share the same weight vector $\mathbf{w}$) but at slightly different positions and scales. C_2 units are therefore selective for the same stimulus as their afferents S_2 units. Yet they are less sensitive to the position and scale of the stimulus within their receptive fields. Receptive field sizes at the C_2 level range between $1.1^o - 3.0^o$. We found that the tuning of model C_2 units (and their invariance properties) to different standard stimuli such as Cartesian and non-Cartesian gratings, two-bar stimuli, and boundary conformation stimuli is compatible with data from V4 [17, 18, 19], (see ref. [3], pp. 28-36 and ref. [20]). SI Fig. 8 shows the details of the model architecture from the S_1 to the C_2 stages.

S_3 and C_3 units

Beyond the S_2 and C_2 stages the same process is iterated once more to increase the complexity of the preferred stimulus at the S_3 level (possibly related to Tanaka's feature columns in TEO, see ref. [20]). For each S_3 unit, the responses of $n_{S_3} = 100$ C_2 units with different selectivities are combined with a tuning operation. The result is an increase of the complexity of the preferred stimulus from the C_2 to the S_3 stages. In the next stage, possibly overlapping between TEO and TE, the complex C_3 units, obtained by pooling S_3 units with the same selectivity at neighboring positions and scales, are also selective to moderately complex features as the S_3 units, but with a larger range of invariance. The pooling parameters of the C_3 units were adjusted so that, at the next stage, units in the S_4 layer exhibit tuning and invariance properties similar to those of the so-called view-tuned cells of AIT [21] (see ref. [16, 3, 20]). The receptive field sizes of the S_3 units are about $1.2^o - 3.2^o$ while the receptive field sizes of the C_3 and S_4 units is about the size of the stimulus ($7^o \times 7^o$ in the present simulation).

S_{2b} and C_{2b} units

In addition to the direct route to AIT (i.e., from S_2 to S_4 through C_3, see Fig. 1) we also implemented bypass routes, that is, direct projections from V2 to TEO (bypassing V4) and from V4 to TE (bypassing TEO) [22]: S_{2b} units combine the response of several retinotopically organized V1-like complex C_1 units at different orientations just like S_2 units. Yet their receptive field is larger (2 to 3 times larger) than the receptive fields of the S_2 units. Importantly, the number of afferents to the S_{2b} units is also larger ($n_{S_{2b}} = 100$ vs. $n_{S_2} = 10$), which results in units which are more selective and more "elaborate" than the S_2 units, yet, less tolerant to deformations. The effect of skipping a stage from C_1 to S_{2b} also results at the C_{2b} level in units that are more selective than other units at a similar level along the hierarchy (C_3 units), and at the same time exhibit a smaller range of invariance to positions and scales. We found that the tuning of the C_{2b} units agree with the read out data from IT [6] (see ref. [3]).

Unsupervised Learning from V2 to IT

The selectivity of the S units, i.e., the set of K_X weight vectors $\mathbf{w^i}$ (see Eq. 1) that are shared across all units within each feature map in layers S_2 and higher (i.e., S_{2b} and S_3), is determined by an unsupervised developmental-like learning stage. During this learning stage the model becomes adapted to the statistics of the natural environment [23] (see ref. [24] for a recent review) and units become tuned to common image-features[2] that occur with high probability in natural images.

Learning in the model is sequential, that is, layers are trained one after another (the entire set of natural images is presented during the training of each individual layers) starting from the bottom with layers S_2 and S_{2b} and then progressing to the top with layer S_3. During this developmental-like learning stage, starting with the S_2 layer, the weights $(\mathbf{w^1}, \ldots, \mathbf{w^{K_{S_k}}})$ of the K_{S_k} feature maps are learned sequentially from 1 to K_{S_k}. At the i^{th} image presentation, one unit at a particular position and scale is selected (at random) from the i^{th} feature-map and is imprinted. That is, the unit stores in its synaptic weights $\mathbf{w^i}$, the current pattern of activity from its afferent inputs (from the previous layer), in response to the part of the natural image that falls within its receptive field. This is done by setting $\mathbf{w^i}$ to be equal to the current pattern of pre-synaptic activity $\mathbf{x}$, such that[3]:

$$\mathbf{w^i} = \mathbf{x}.$$

As a result the image patch $\mathbf{x}$ that falls within the receptive field of the unit $\mathbf{w^i}$ becomes its preferred stimulus. Note that units in higher layers are thus tuned to larger patches. During this learning stage, we also assume that the image moves (shifts and looms) so that the selectivity of the unit that was just imprinted is generalized to units in the same feature map across scales and positions [4]. After this imprinting stage the feature map i is mature and the synaptic weight $\mathbf{w^i}$ of the units within the map is fixed. Learning all K_{S_k} unit types within the S_k layer thus requires K_{S_k} image presentations. The database of images we used contains a large variety of natural images collected from the web (including landscapes, street scenes, animals, etc).

As a result of this new learning stage, the architecture of Fig. 1 contains a total of 10^7 tuned units. At the top of the hierarchy, the classification units rely on a dictionary of $6,000$ units tuned to image features with different levels of selectivities and invariances. This is 2-3 orders of magnitude larger than the number of features used by both biological models as well as current computer vision systems (e.g., ref. [27]) that typically rely on 10-100 features.

The d' sensitivity measure

The d' sensitivity measure [36] is a performance measure which, for each observer, combines both the hit rate H, that is, the proportion of animal images correctly classified by the observer, and the false alarm rate F, that is, the proportion of non-animal images incorrectly classified by the observer into one single standardized score. The mathematical form of the d' measure is

$$d' = Z(H) - Z(F),$$

where Z corresponds to the inverse of the normal distribution function.

[2] The resulting hierarchy of unit selectivities in the model is related to other approaches such as component-based [25], part-based [26] or fragment-based approaches [27] in computer vision. This is also sometime referred to as "bags of features" in computer vision or "unbound features" [28, 29, 30] in cognitive science.

[3] A biophysical implementation of this rule would involve mechanisms such as LTP [31, 32, 33, 34].

[4] In the present version of the model this is done by simply "tiling" units. During biological development of the circuitry, this could involve a generalized Hebbian rule [35] (T. Masquelier, T. Serre, S. Thorpe & T. Poggio, in prep).

On interrupting recurrent processing with the mask

The effect of a backward mask – as used in the psychophysics described here – on visual processing remains a matter of debate (see ref. [37] for a recent overview). A well accepted theory is the "interruption theory" that has been in fact corroborated by physiological studies in V1 [38, 39, 40], IT [41, 42], STS [43] and FEF [44] (see also ref. [37, 45, 46] for recent reviews). The assumption is that the visual system processes stimuli sequentially (in a pipeline-like architecture): when a new stimulus (the mask) is piped in it interrupts the processing of the previous stimulus (the target image). Importantly these physiological studies have shown that the mask tend to leave intact the early part of the neural response (corresponding to the original feedforward sweep from bottom-up inputs) while disrupting the late part of the neural response (modulated by feedbacks from higher areas, see ref. [40] for instance). This suggests that a backward mask can be used as a tool to isolate between feedforward-dominated vs. recurrent processing (i.e., incorporating both feedforward and feedback loops).

Under the "interruption theory", whether or not specific feedback loops (say between PFC and V4 or IT and V4) participate in the overall processing is determined by the delay between the stimulus and the mask (i.e., the SOA). If the delay Δ taken by the visual signal to travel from stage A (e.g., V4) to stage B (e.g., V1) and back to stage A is longer than the SOA, this specific back-projection (from B to A) will not influence the processing of the target as the mask signal will reach stage A before the target signal has had time to reach A back from B.

Based on estimates of conduction delays (see SI Fig. 7), extrapolated from monkey [47, 48] to human (S. Thorpe, Personal communication), we think that in all our experiments, a SOA of 50 ms is likely to be the longest SOA before significant feedback loops become active [5], for instance, between IT and V4 (see SI Fig. 7, orange arrows, $\Delta \sim$ 40-60 ms). Importantly such an SOA should exclude major top-down effects, for instance between IT and V1 ($\Delta \sim$ 80-120 ms), while leaving enough time for signal integration at the neural level[6]. This estimate seems in good agreement with results from a Transcranial Magnetic Stimulation (TMS) experiment [51] that has shown a disruption of the feedforward sweep [45] for pulses applied between 30 ms and 50 ms after stimulus onset.[7] It is thus quite interesting that the model matches human performance for an SOA of 50 ms, but underperforms it for longer SOAs. One of the possible explanations is that this is due to back-projections which are not included in the present, purely feedforward model of Fig. 1. Insightful discussions on the role of the back-projections in visual processing can be found in [45, 52].

Benchmarking the database

To ensure that the animal vs. non-animal discrimination task cannot be performed based solely on low level features, we evaluated several benchmark computer vision systems on the database of stimuli. This includes two simple systems (one based on the mean luminance of the images and another based on the pixel values – similar to a retina – directly passed to a single template SVM classifier). We also ran two standard computer vision systems that were previously compared to human observers in rapid categorization tasks: a texton-based system [53] and a global feature-based system [54]. Finally, to evaluate the contribution of intermediate model layers, we used the activity of the C_1 layer (corresponding to complex cells in V1) that we passed to a linear SVM classifier directly. Details about the implementations of these benchmark systems can be found online on our supplementary web material [1].

The performance of the different approaches is summarized in SI Table 2. The simplest systems (mean luminance and single template SVM classifier) perform very poorly, suggesting that the task is non-trivial. While the computer vision systems [53, 54] as well as the model C_1 layer perform better, their level of performance remains lower than the level of performance of the human observers and the model.

Altogether the comparative superiority of the model over the benchmark systems suggest the need for a representation based on units with different levels of complexity and invariance as in the architecture of Fig. 1 (see main manuscript). Consistent with the results reported here, an independent study (see ref. [4], pp. 42–50) found a gradual improvement (using layers in the model from bottom to top) in reading out several object categories (at different positions and scales) from various model layers (see also SI Fig. 4 for a comparison between different layers of the model on the animal / non-animal categorization task).

We would like to thank Stan Bileschi for writting the software and running the texton algorithm in section *Benchmarking the database*.

1. Serre, T, Oliva, A, & Poggio, T. (2007) Supplementary web material. http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07/index.htm.
2. Poggio, T & Smale, S. (2003) *Notices of the american Mathematical Society (AMS)* **50**.
3. Serre, T, Kouh., M, Cadieu, C, Knoblich, U, Kreiman, G, & Poggio, T. (2005) *MIT AI Memo 2005-036 / CBCL Memo 259*.
4. Serre, T, Kouh., M, Cadieu, C, Knoblich, U, Kreiman, G, & Poggio, T. (2005) A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex, (MIT, Cambridge, MA), AI Memo 2005-036 / CBCL Memo 259.
5. Perrett, D, Hietanen, J, Oram, M, & Benson, P. (1992) *Philos. Trans. Roy. Soc. B* **335**, 23–30.
6. Hung, C, Kreiman, G, Poggio, T, & DiCarlo, J. (2005) *Science* **310**, 863–866.
7. Smith, E & Lewicki, M. (2006) *Nature* **439**, 978–982.
8. Grossberg, S. (1973) *Studies in Applied Mathematics* **52**, 213–257.
9. VanRullen, R, Guyonneau, R, & Thorpe, S. (2005) *Trends in Neurosci.* **28**.
10. Schiller, P. H, Finlay, B. L, & Volman, S. F. (1976) *J. Neurophysiol.* **39**, 1288–1319.
11. Hubel, D. H & Wiesel, T. N. (1965) *J. Neurophys.* **28**, 229–289.
12. Serre, T & Riesenhuber, M. (2004) Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex, (MIT, Cambridge, MA), AI Memo 2004-017 / CBCL Memo 239.
13. Jones, J. P & Palmer, L. A. (1987) *J. Neurophys.* **58**, 1233–1258.
14. Hubel, D. H & Wiesel, T. N. (1968) *J. Phys.* **195**, 215–243.
15. DeValois, R, Albrecht, D, & Thorell, L. (1982) *Vis. Res.* **22**, 545–559.
16. Serre, T & Riesenhuber, M. (2004) *MIT AI Memo 2004-017 / CBCL Memo 239*.
17. Gallant, J, Connor, C, Rakshit, S, Lewis, J, & Essen, D. V. (1996) *J. Neurophys.* **76**, 2718–2739.
18. Pasupathy, A & Connor, C. E. (2001) *J. Neurophys.* **86**, 2505–2519.
19. Reynolds, J. H, Chelazzi, L, & Desimone, R. (1999) *J. Neurosci.* **19**, 1736–1753.
20. Serre, T. (2006) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge, MA).
21. Logothetis, N. K, Pauls, J, & Poggio, T. (1995) *Curr. Biol.* **5**, 552–563.
22. Nakamura, H, Gattass, R, Desimone, R, & Ungerleider, L. G. (1993) *J. Neurosci.* **13**, 3681–3691.
23. Attneave, F. (1954) *Psychol. Rev.* **61**, 183–193.
24. Simoncelli, E & Olshausen, B. (2001) *Ann. Rev. Neurosci.* **24**, 1193–1216.

[5] Note that for such an SOA, local feedback loops green arrows in SI Fig. 7) are likely to be already active ($\Delta <$20-30 ms), see ref. [49, 50].

[6] The mask is likely to interrupt the maintained response of IT neurons but not to alter their initial selective response [41, 42]. According to an independent study [6] this would provide significantly more time than needed ($\gg$ 12.5 ms) to permit robust recognition in "reading out" from monkey IT neurons.

[7] The same experiment [51] also demonstrated blockade of perception by pulses applied between 80-120 ms, presumably corresponding to recurrent processing [45] by the back-projections.

25. Heisele, B, Serre, T, Pontil, M, Vetter, T, & Poggio, T. (2002) *Advances in Neural Information Processing Systems* **14**, 1239–1245.
26. Fei-Fei, L, Fergus, R, & Perona, P. (2004) *Proc. IEEE CVPR, Workshop on Generative-Model Based Vision*.
27. Ullman, S, Vidal-Naquet, M, & Sali, E. (2002) *Nat. Neurosci.* **5**, 682–687.
28. Treisman, A. M & Gelade, G. (1980) *Cog. Psych.* **12**, 97–136.
29. Evans, K & Treisman, A. (2005) *J. Exp. Psych.: Hum. Percept. Perf.* **31**, 1476–1492.
30. Wolfe, J & Bennett, S. (1997) *Vis. Res.* **37**, 25–44.
31. Markram, H, Lübke, J, Frotscher, M, & Sakmann, B. (1997) *Science* **275**, 213–215.
32. Bi, G & Poo, M. (1998) *J. Neurosci.* **18**, 10464–10472.
33. Abarbanel, H, Huerta, R, & Rabinovich, M. (2002) *Proc. Nat. Acad. Sci. USA* **99**, 10132–10137.
34. van Rossum, M, Bi, G, & Turrigiano, G. (2000) *J. Neurosci.* **20**, 8812–8821.
35. Földiák, P. (1991) *Neural Comp.* **3**, 194–200.
36. Macmillan, N. A & Creelman, C. D. (1991) *Detection Theory: A User's Guide*. (Cambridge University Press).
37. Breitmeyer, B & Ogmen, H. (2006) *Visual Masking: Time Slices through Conscious and Unconscious Vision*. (Oxford University Press).
38. Bridgeman, B. (1980) *Brain Res.* **196**, 347–364.
39. Macknik, S & Livingstone, M. (1998) *Nat. Neurosci.* **1**, 144–149.
40. Lamme, V, Zipser, K, & Spekreijse, H. (2002) *J. Cogn. Neurosci.* **14**, 1044–1053.
41. Kovács, G, Vogels, R, & Orban, G. (1995) *Proc. Nat. Acad. Sci. USA* **92**, 5587–5591.
42. Rolls, E, Tovee, M, & Panzeri, S. (1999) *J. Cogn. Neurosci.* **11**, 300–311.
43. Keysers, C, Xiao, D. K, Földiák, P, & Perrett, D. I. (2001) *J. Cogn. Neurosci.* **13**, 90–101.
44. Thompson, K & Schall, J. (1999) *Nat. Neurosci.* **2**, 283–288.
45. Lamme, V & Roelfsema, P. (2000) *Trends in Neurosci.* **23**, 571–579.
46. Enns, J & Lollo, V. D. (2000) *Trends in Cogn. Sci.* **4**, 345–351.
47. Nowak, L & Bullier, J. (1997) *Extrastriate visual cortex in primates*. (New York: Plenum Press) Vol. 12, pp. 205–241.
48. Thorpe, S & Fabre-Thorpe, M. (2001) *Science* **291**, 260–263.
49. Knierim, J & van Essen, D. (1992) *J. Neurophys.* **67**, 961–p80.
50. Zhou, H, Friedman, H. S, & von der Heydt, R. (2000) *J. Neurosci.* **20**, 6594–6611.
51. Corthout, E, Uttl, B, Walsh, V, Hallett, M, & Cowey, A. (1999) *Neuroreport*.
52. Hochstein, S & Ahissar, M. (2002) *Neuron* **36**, 791–804.
53. Renninger, L & Malik, J. (2004) *Vis. Res.* **44**, 2301–2311.
54. Torralba, A & Oliva, A. (2003) *Network: computation in neural systems* **14**, 391–412.
55. Kobatake, E, Wang, G, & Tanaka, K. (1998) *J. Neurophys.* **80**, 324–330.