

A feedforward theory of visual cortex accounts for human performance in rapid categorization

Thomas Serre^{® 1,2,3,4}, Aude Oliva^{3,4}, Tomaso Poggio^{1,2,3,4}

¹ *McGovern Institute for Brain Research*

² *Center for Biological and Computational Learning*

³ *Department of Brain and Cognitive Sciences*

⁴ *Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*

® *To whom correspondence should be addressed, e-mail: serre@mit.edu*

Primates are remarkably good at recognizing objects in cluttered natural images. The level of performance of the primate visual system and its robustness to image variability have remained unchallenged by the best computer vision systems despite decades of engineering effort. We developed a new model of the feedforward path of the ventral stream in primate visual cortex that incorporates many anatomical and physiological constraints. Its key property – in addition to supervised learning from IT cortex to higher areas – is an unsupervised learning stage that creates from natural images a large generic dictionary of tuned units from V2 to IT useful for different recognition tasks. Remarkably, these model units exhibit tuning properties consistent with the known physiology of the main visual areas. Here we report that the model can predict both the level and the pattern of performance achieved by humans on a difficult animal vs. non-animal rapid categorization task. The high performance of a feedforward, hierarchical model compared with existing computer vision systems and with human vision supports a theoretical framework for understanding properties of single neurons and cortical areas in the context of high level visual functions while suggesting a novel architecture for computer vision systems.

Object recognition in cortex is mediated by the ventral visual pathway running from primary visual cortex¹, V1, through extrastriate visual areas V2 and V4 to inferotemporal cortex²⁻⁴, IT (comprising PIT and AIT), and then to prefrontal cortex (PFC) which is involved in linking perception to memory and action. It is well known that recognition is possible for scenes viewed in rapid visual presentation that do not allow sufficient time for eye movements⁵⁻¹⁰ and in the near-absence of attention¹¹. The hypothesis that the basic processing of information is feedforward is supported most directly by the short times required for a selective response to appear in IT cells¹². Very recent data¹³ convincingly show that the activity of small neuronal populations in monkey IT, over very short time intervals (as small as 12.5 ms) and only about 100 ms after stimulus onset, contains surprisingly accurate and robust information supporting a variety of recognition tasks. While this does not rule out the use of

local feedback loops within an area, it does suggest that a core hierarchical feedforward architecture may be a reasonable starting point for a theory of visual cortex, aiming to explain “immediate recognition” – the initial phase of recognition before eye movements and high-level processes can play a role.

We here describe a new quantitative model that accounts for the circuits and computations of the feedforward path of the ventral stream of visual cortex. This model is consistent with a general theory of visual processing that extends the hierarchical model of Hubel & Wiesel from primary to extrastriate visual areas. This theory builds upon several existing neurobiological models¹⁴⁻²¹ and conceptual proposals^{1,2,22}. An implementation of the theory is sketched in Fig. 1, with full details provided in Methods (see SI, section A.1). It follows the basic architecture and relies on the same basic two operations of¹⁷.

Its key new aspect is the learning of a generic dictionary of shape-components from V2 to IT, which provides a rich representation to task-specific categorization circuits in higher brain areas. Importantly, the hierarchical architecture builds progressively more invariance to position and scale while preserving the selectivity of the units. This vocabulary of tuned units is learned from natural images during a developmental-like, unsupervised learning stage in which each unit in the S2, S2b and S3 layers becomes tuned to a different patch of a natural image (randomly chosen, see SI, section A.1.2). The resulting dictionary is generic and universal, in the sense that it can support several different recognition tasks and, in particular, the recognition of many different object categories²³ (other models, see¹⁸ for instance, learn features that are specific to the task, thus requiring learning an entire new set of features for every new category to be learned). For the “mature” model to learn a new categorization task, only the task-specific circuits at the top level in the model, possibly corresponding to categorization units in PFC²⁴, have to be trained from a small set of labeled examples and in a task specific manner (see SI, section A.1.3).

The model of Fig. 1 is characterized by two major advances with respect to previous attempts. First the new model is significantly closer to the anatomy and the physiology of visual cortex with more layers (reflecting PIT as well as V4) and with a looser hierarchy (reflecting the bypass connections from V2 to PIT and V4 to AIT²⁵). It is indeed qualitatively and quantitatively consistent with (and in some cases actually predicts, see²³) several properties of cells in V1, V2, V4, and IT, PFC²⁴ as well as several fMRI and psychophysical data²⁶. For instance, the model predicts²³, at the C1 and C2 levels respectively, the max-like behavior of a subclass of complex cells in V1²⁷ and V4²⁸. It also agrees²³ with other data in V4²⁹ about the response of neurons to combinations of simple two-bar stimuli (within the receptive field of the S2 units) and some of the C2 units in the model show a tuning for boundary conformations which is consistent with recordings from V4³⁰ (see Cadieu, Kouh, Pasupathy, Connor et al, in prep). Read-out from C2b units in the model predicted²³ recent read-out experiments in IT¹³, showing very similar selectivity and invariance for the same set of stimuli.

Second, not only can the model duplicate the tuning properties of neurons in various brain areas when probed with artificial stimuli, but, it can also handle the recognition of objects in the real-world: the model performs much better than other less detailed biologically motivated feedforward models¹⁴⁻²¹ — and performs at least as well as state-of-the-art computer vision systems which do not have any relation with the anatomy and the physiology of the ventral stream (see SI, section B). Key to the good recognition performance of the model is the large number of tuned units across the hierarchical architecture of the model which are learned from natural images and represent a redundant dictionary of fragment-like features^{18,31,32} that span a range of selectivities and invariances. As a result of this new learning stage, the architecture of Fig. 1 contains a total of ~10 million tuned units (see SI, section C.1). At the top, 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 state-of-the-art computer vision systems (see SI, Table S3) that typically rely on 10-100 features. In addition, the model is remarkably robust

to parameter values, detailed wiring and even exact form of the two basic operations and of the learning rule (see ²³ and SI).

Because this new feedforward model agrees with many physiological data while performing surprisingly well in the recognition of natural images it is natural to ask how well it may predict human performance in complex object recognition tasks. Of course, as a feedforward model of the ventral stream pathway, the architecture of Fig. 1 cannot account for our everyday vision which involves eye movements and top-down effects, which are mediated by higher brain centers and the extensive anatomical back-projections found throughout visual cortex and not implemented in the present feedforward model.

A natural paradigm for comparing the performance of human observers in an object recognition task to that of a feedforward model of visual processing is ultra-rapid categorization. A well-established experiment is an animal vs. non-animal recognition task ^{6-9,33}. Animals in natural scenes constitute a challenging class of stimuli due to large variations in shape, pose, size, texture, and position in the scene (see SI, Table S2 for the performance of several benchmark systems). To vary the difficulty of the task, we used four sets of balanced image categories (150 animals and 150 matching distractors, see SI, section A.2.4), each corresponding to a particular viewing-distance from the camera, from an animal head to a small animal or groups of animals in cluttered natural backgrounds (i.e. “head”, “close-body”, “medium-body” and “far-body” categories, see Fig. 2a, and SI, section A.2.3). We used a backward masking protocol (1/f noise image with a duration of 80 ms, see Fig. 2b) with a “long” 50 ms stimulus onset asynchrony (50 ms SOA corresponding to a 20 ms stimulus presentation followed by a 30 ms inter-stimulus interval, see SI, section A.2.5) to give us close to ceiling performance ⁹ while trying to block significant top-down effects through the back-projections (see later and SI, section C.2). In the model, processing by the units (the nodes of the graph in Fig. 1) is approximated as essentially instantaneous, with the processing time taken

by synaptic latencies and conduction delays (see²³ and SI, section C.2). Thus the model does not see the mask.

A comparison between the performance of human observers ($n = 24$, 50 ms SOA) and the feedforward model in the animal classification task is shown in Fig. 3a. Performance is measured by the d' (other accuracy measures such as error rates or hits gave similar results, see SI, section D.3) which is a sensitivity measure that combines both the hit and false-alarm rates of each observer into one standardized score (see SI, section A.2). The task-specific circuits of the model were trained for the animal vs. non-animal categorization task in a supervised way using a random split procedure (see SI, section A.2.2 and A.2.5) on the entire database of stimuli (i.e., in a given run, half the images were selected at random for training and the other half were used for testing the model). Human observers and the model behave similarly: across all four animal categories, their levels of performance do not show significant differences (with overall correct 80% for human observers and 82% for the model). It should be noted that no single model parameter was adjusted to fit the human data (all parameters were adjusted before the experiment to account for physiology data only). The accuracy of the human observers is well within the range of data previously obtained with go/no-go tasks on similar tasks^{6,8,9}.

Additionally, both the model and human observers tend to produce similar responses (both correct and incorrect, see Fig. 3). We measured quantitatively the agreement between human observers and the model on individual images. For each image in the database, we computed the percentage of observers (black number above each thumbnail) who classified it as an animal (irrespective of whether the image contains an animal or not). For the model, we computed the percentage of times the model (green number) classified each image as an animal for each of the random runs (during each run, the model is trained and tested on a different set of images and therefore, across several runs, the same test image may be classified differently by the model.). A percentage of 100% (50%) means that all (half) the observers (either human observers or random runs of the model) classified this image as an animal. The overall image-

by-image correlation between the model and human observers is high (specifically 0.71, 0.84, 0.71 and 0.60 for heads, close-body, medium-body and far-body respectively, with $p < 0.01$). Together with the results of a "lesion study" performed on the model (see SI, Fig. S6), the data suggest that the overall set of features (of "moderate complexity") from V2 to V4 and PIT underlies such a performance in this task.

To further challenge the model, we measured the effect of image rotation (90° and 180°) on performance. Recent behavioral studies^{33,34} suggested that the animal categorization task can be performed very well by human observers on rotated images. Can the model predict human behavior in this situation? Fig. 4 shows that the model (right) and human observers (left) show a similar pattern of performance and are similarly robust to image rotation. The robustness of the model is particularly remarkable as it was not re-trained before being tested on the rotated images. It is likely due to the fact that an image patch of a rotated animal is more similar to an image patch of an upright animal than to a non-animal.

Finally, we replicated previous psychophysical results⁹ to test 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 (see Fig. S5 and SI, section D.1). For SOAs longer than 80 ms, human observers outperform the model.

It remains an open question whether the somewhat better performance of humans for SOAs longer than 80 ms is due to feedback effects mediated by the back-projections. Previous physiological studies have suggested that a backward mask can interrupt visual processing and

block back-projections³⁵⁻³⁷ (see also³⁸). Under this assumption (see SI, section C.2), we estimate from physiology data on response latencies that the major effects of back-projections may appear for SOAs around 40-60 ms. The model indeed mimics human-level performance for the 50 ms condition. 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. An alternative hypothesis is that a) back-projections do not have any role in categorization even for very long SOAs and that b) the feedforward model could still be somewhat improved to attain human performance in the absence of a mask.

Our main result is in any case that a relatively simple hierarchical architecture, reflecting the known physiology and anatomy of visual cortex, achieves a high level of correlation with humans – and comparable accuracy – on a difficult (but rapid) recognition task.

Acknowledgements. We are grateful to C. Cadieu, B. Desimone, C. Koch, M. Riesenhuber, D. Perrett, U. Knoblich, M. Kouh, G. Kreiman, S. Thorpe, A. Torralba and J. Wolfe, for comments and fruitful discussions related to this work. We would also like to thank S. Das, J. Dicarlo, D. Ferster, M. Giese, M. Greene, E. Meyers, E. Miller, P. Sinha, C. Tomasi and R. VanRullen for comments on this manuscript. This research was sponsored by grants from NIH, DARPA, ONR and the National Science Foundation. Additional support was provided by Eastman Kodak Company, Daimler Chrysler, Honda Research Institute, NEC Fund, Siemens Corporate Research, Toyota, Sony and the McDermott chair (T.P.).

Figure Legends

Fig. 1. Tentative mapping between the ventral stream in the primate visual system (left) and the functional primitives of the feedforward model (right). The model accounts for a set of basic facts about the cortical mechanisms of recognition that have been established over the last decades: From V1 to IT, there is an increase in invariance to position and scale^{1,2,4,17,39} and, in parallel, an increase in the size of the receptive fields^{2,4} as well as in the complexity of the optimal stimuli for the neurons^{2,3,40}. Finally adult plasticity and learning are probably present at all stages, and certainly at the level of IT³⁹ and PFC. The theory assumes that one of the main functions of the ventral stream – just a part of visual cortex – is to achieve a trade-off between selectivity and invariance within a hierarchical architecture. As in¹⁷, stages of “simple” (S) units with Gaussian tuning (plain circles and arrows), are loosely interleaved with layers of “complex” (C) units (dotted circles and arrows), which perform a max-like operation on their inputs and provide invariance to position and scale (pooling over scales is not shown in the figure). The tuning of the S2, S2b and S3 units (corresponding to V2, V4 and PIT, respectively) is determined here by a prior developmental-like unsupervised learning stage (see SI, section A.1.2). Learning of the tuning of the S4 units and of the synaptic weights from S4 to the top classification units is the only task-dependent, supervised learning stage. The main route to IT is denoted with black arrows while the bypass route²⁵ is denoted with blue arrows 1 (see section 1.1.1 pp 6 for details). The total number of units in the model simulated in this paper is in the order of 10 million. Colors indicate the correspondence between model layers and cortical areas. The table on the right provides a summary of the main properties of the units at the different levels of the model. Note that the model is a simplification and only accounts for the ventral stream of visual cortex. Of course other cortical areas (e.g., in the dorsal stream) as well as non-cortical structures (e.g., basal ganglia) are likely to play a role in the process of object recognition. The diagram on the left is modified from Van Essen & Ungerleider⁴¹ (with permission by the authors).

Fig. 2. Animal vs. Non-animal categorization task.

a) The four (balanced) classes of stimuli. Animal images (a subset of the image database used in ⁶) were manually arranged into four groups (150 images each) based on the animal-distance from the camera: head (close-up), close-body (animal body occupying the whole image), medium-body (animal in scene context) and far-body (small animal or groups of animals). Each of the four classes corresponds to different animal sizes and, probably through the different amount of clutter relative to the object size, modulates the task difficulty. A set of matching distractors (300 each from natural and artificial scenes, see SI, section A.2.3) was selected, so as to prevent human observers and the computational model from relying on low-level cues (see SI, section A.2.4).

b) Schematic of the task. A stimulus (gray-level image) is flashed for 20 ms, followed by a blank screen for 30 ms (i.e., SOA of 50 ms) and followed by a mask for 80 ms. Subjects ended the trial with a yes/no answer by pressing one of two keys.

Fig. 3. Comparison between the model and human observers.

a) Model vs. human-level accuracy. Human observers and the model exhibit a very similar pattern of performance (measured with d' measure, see SI, section A.2.1). Error bars indicate the standard errors for the model (computed over $n = 20$ random runs) and for human observers (computed over $n = 24$ observers).

Examples of classifications by the model and human observers. Common false alarms (b) and misses (c) for the model and human observers. Examples of animal images for which the agreement between the model and human observers is (d) poor and (e) good. The percentages above each thumbnail correspond to the number of times the image was classified as animal by the model (green number) or by human observers (black number, see text for details). Part of the discrepancy between the model and human observers is likely to be due to the relatively small set of examples used to train the model (300 animal and 300 non-animal images).

Fig. 4. Effect of image orientation.

Comparison between the performance (d') of the human observers (left, $n = 14$) and the model (right) in three experimental conditions: upright, 90° rotation and inverted (180°) image presentations. Human observers and the model are similarly robust to image rotations.

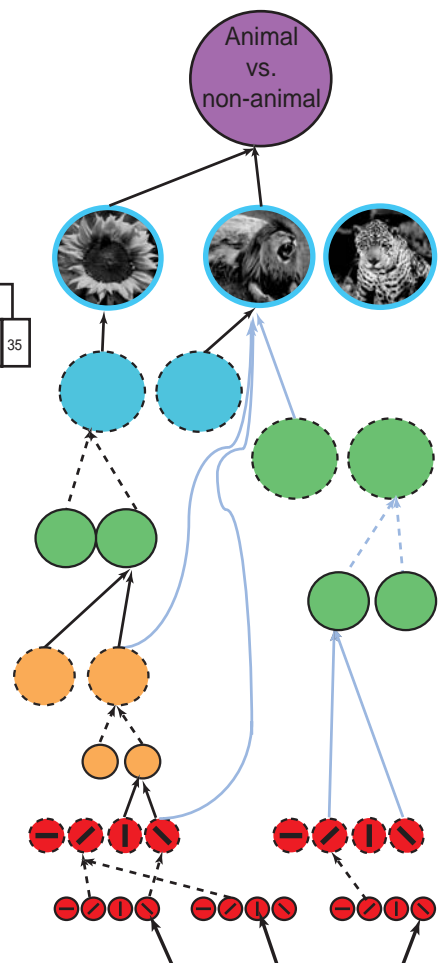
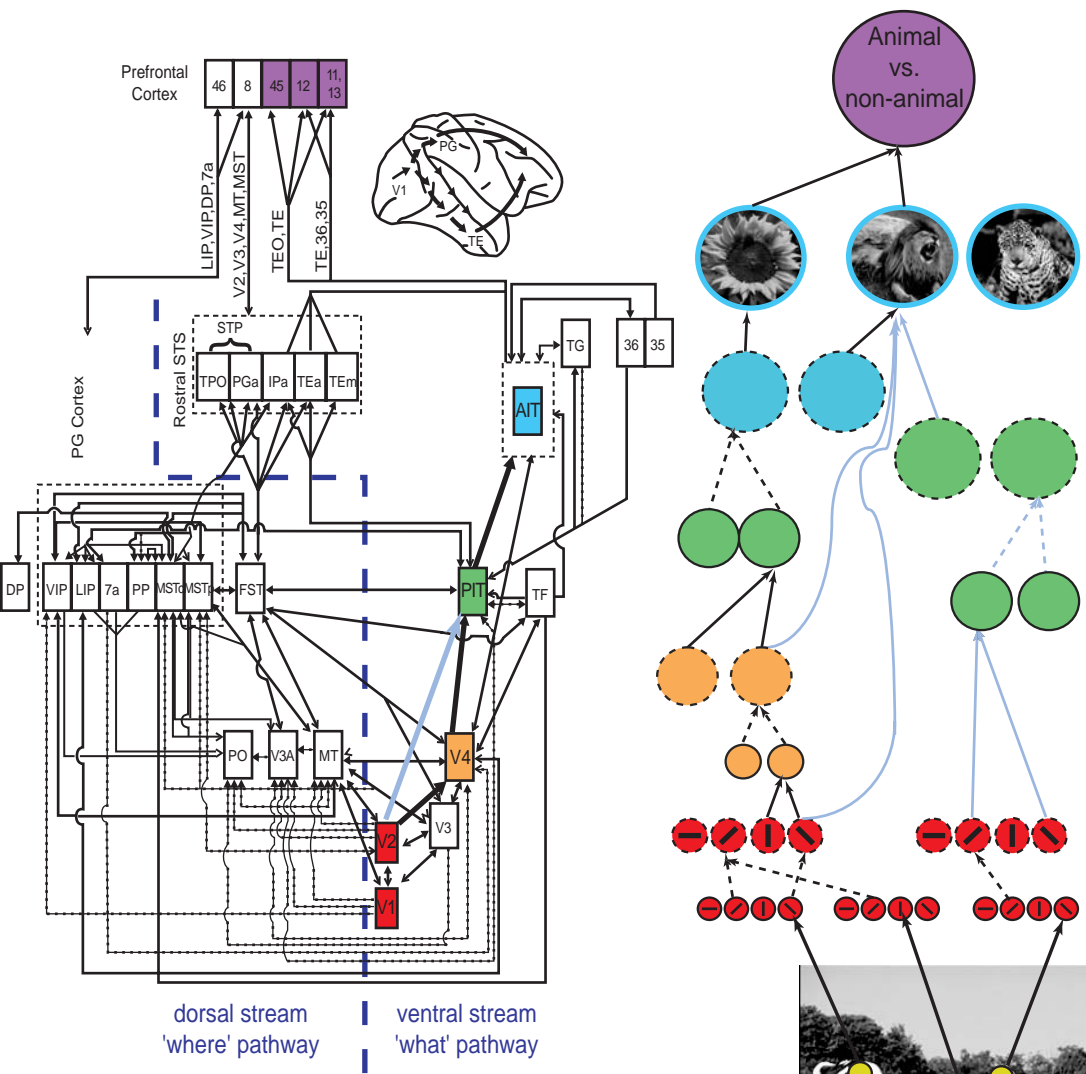
Reference List

- ¹ D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *J. Physiol* **195**(1), 215 (1968).
- ² D. I. Perrett and M. W. and Oram, "Neurophysiology of shape processing," *Img. Vis. Comput.* **11**, 317 (1993).
- ³ E. Kobatake and K. Tanaka, "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex," *J. Neurophysiol.* **71**(3), 856 (1994).
- ⁴ K. Tanaka, "Inferotemporal cortex and object vision," *Annu. Rev. Neurosci.* **19**, 109 (1996).
- ⁵ M. C. Potter, "Meaning in visual search," *Science* **187**(4180), 965 (1975).
- ⁶ S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature* **381**(6582), 520 (1996).
- ⁷ S. J. Thorpe, *et al.*, "Detection of animals in natural images using far peripheral vision," *Eur. J. Neurosci.* **14**(5), 869 (2001).
- ⁸ R. VanRullen and C. Koch, "Visual selective behavior can be triggered by a feed-forward process," *J. Cogn Neurosci.* **15**(2), 209 (2003).
- ⁹ N. Bacon-Mace, *et al.*, "The time course of visual processing: backward masking and natural scene categorisation," *Vision Res.* **45**(11), 1459 (2005).
- ¹⁰ H. Kirchner and S. J. Thorpe, "Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited," **46**, 1762 (2005).
- ¹¹ F. F. Li, *et al.*, "Rapid natural scene categorization in the near absence of attention," *Proc. Natl. Acad. Sci. U. S. A* **99**(14), 9596 (2002).
- ¹² D. I. Perrett, *et al.*, "Organization and functions of cells responsive to faces in the temporal cortex," *Philos. Trans. R. Soc. Lond B Biol. Sci.* **335**(1273), 23 (1992).

- ¹³ C Hung, *et al.*, "Fast Readout of Object Identity from Macaque Inferior Temporal Cortex," *Science* **310**(5749), 863 (2005).
- ¹⁴ K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cyb.* **36**, 193 (1980).
- ¹⁵ G. Wallis and E. T. Rolls, "Invariant face and object recognition in the visual system," *Prog. Neurobiol.* **51**(2), 167 (1997).
- ¹⁶ B. W. Mel, "SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition," *Neural Comput.* **9**(4), 777 (1997).
- ¹⁷ M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.* **2**(11), 1019 (1999).
- ¹⁸ S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual features of intermediate complexity and their use in classification," *Nat. Neurosci.* **5**(7), 682 (2002).
- ¹⁹ S. Thorpe, "Ultra-Rapid Scene Categorisation with a Wave of Spikes," , 1 (2002).
- ²⁰ Y. Amit and M. Mascaró, "An integrated network for invariant visual detection and recognition," *Vision Res.* **43**(19), 2073 (2003).
- ²¹ H. Wersing and E. Korner, "Learning optimized features for hierarchical models of invariant object recognition," *Neural Comput.* **15**(7), 1559 (2003).
- ²² S. Hochstein and M. Ahissar, "View from the top: hierarchies and reverse hierarchies in the visual system," *Neuron* **36**(5), 791 (2002).
- ²³ T. Serre, *et al.*, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in visual cortex," (2005).
- ²⁴ D. J. Freedman, *et al.*, "Categorical representation of visual stimuli in the primate prefrontal cortex," *Science* **291**(5502), 312 (2001).
- ²⁵ H. Nakamura, *et al.*, "The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques," *J. Neurosci.* **13**(9), 3681 (1993).
- ²⁶ M. Riesenhuber, *et al.*, "Face processing in humans is compatible with a simple shape-based model of vision," *Proc. Biol. Sci.* **271 Suppl 6**, S448-S450 (2004).
- ²⁷ I. Lampl, *et al.*, "Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex," *J. Neurophysiol.* **92**(5), 2704 (2004).

- ²⁸ T. J. Gawne and J. M. Martin, "Responses of primate visual cortical V4 neurons to simultaneously presented stimuli," *J. Neurophysiol.* **88**(3), 1128 (2002).
- ²⁹ J. H. Reynolds, L. Chelazzi, and R. Desimone, "Competitive mechanisms subserve attention in macaque areas V2 and V4," *J. Neurosci.* **19**(5), 1736 (1999).
- ³⁰ A. Pasupathy and C. E. Connor, "Shape representation in area V4: position-specific tuning for boundary conformation," *J. Neurophysiol.* **86**(5), 2505 (2001).
- ³¹ K. K. Evans and A. Treisman, "Perception of objects in natural scenes, is it really attention-free?," **31**(6), 1476 (2005).
- ³² J. M. Wolfe and S. C. Bennett, "Preattentive object files: shapeless bundles of basic features," *Vision Res.* **37**(1), 25 (1997).
- ³³ G. A. Rousselet, M. J. Mace, and M. Fabre-Thorpe, "Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes," *J. Vis.* **3**(6), 440 (2003).
- ³⁴ R. Guyonneau, H. Kirchner, and S. J. Thorpe, "Animals roll around the clock: The rotation invariance of ultra-rapid visual processing," (2005).
- ³⁵ G. Kovacs, R. Vogels, and G. A. Orban, "Cortical correlate of pattern backward masking," *Proc. Natl. Acad. Sci. U. S. A* **92**(12), 5587 (1995).
- ³⁶ E. T. Rolls, M. J. Tovee, and S. Panzeri, "The neurophysiology of backward visual masking: information analysis," *J. Cogn Neurosci.* **11**(3), 300 (1999).
- ³⁷ C. Keysers, *et al.*, "The speed of sight," *J. Cogn Neurosci.* **13**(1), 90 (2001).
- ³⁸ V. A. Lamme and P. R. Roelfsema, "The distinct modes of vision offered by feedforward and recurrent processing," *Trends Neurosci.* **23**(11), 571 (2000).
- ³⁹ N. K. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Curr. Biol.* **5**(5), 552 (1995).
- ⁴⁰ R. Desimone, "Face-selective cells in the temporal cortex of monkeys," *J. Cogn. Neurosci.* **3**, 1 (1991).
- ⁴¹ C. G. Gross, *brain vision and memory: tales in the history of neuroscience* (MIT Press, 1998).

Poggio, Fig. 1

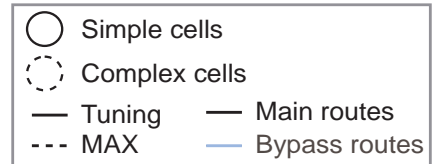


Model layers	Corresponding brain area (tentative)	RF sizes	Num. units	Num. subunits
classifier	PFC		10^0	
S4	AIT	7°	10^2	6,000
C3	PIT - AIT	7°	10^3	
C2b	PIT	7°	10^3	
S3	PIT	$1.2^\circ - 3.2^\circ$	10^4	100
S2b	V4 - PIT	$0.9^\circ - 4.4^\circ$	10^7	100
C2	V4	$1.1^\circ - 3.0^\circ$	10^5	
S2	V2 - V4	$0.6^\circ - 2.4^\circ$	10^7	10
C1	V1 - V2	$0.4^\circ - 1.6^\circ$	10^4	
S1	V1 - V2	$0.2^\circ - 1.1^\circ$	10^6	

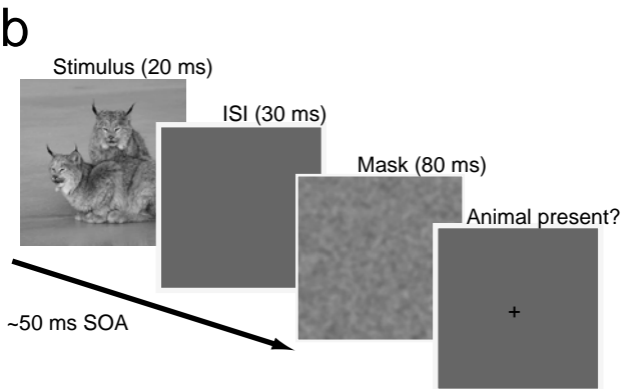
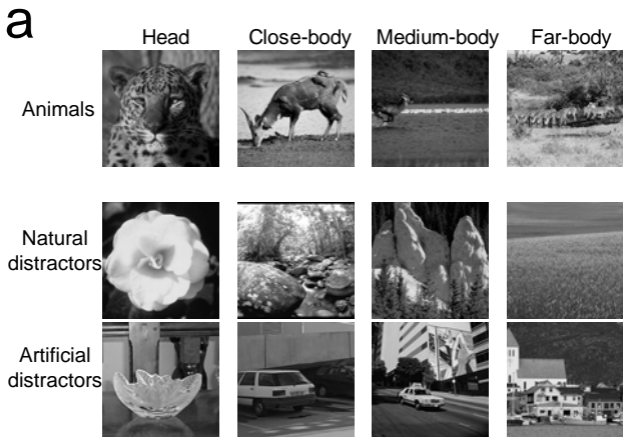
Supervised task-dependent learning

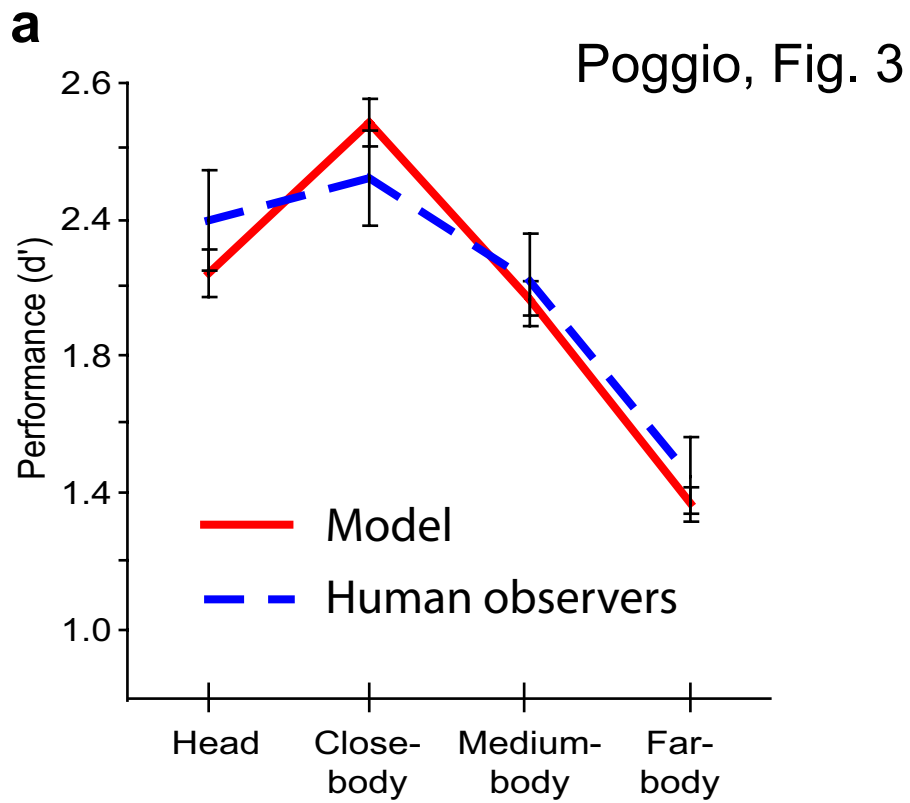
Unsupervised task-independent learning

Increase in complexity (number of subunits), RF size and invariance

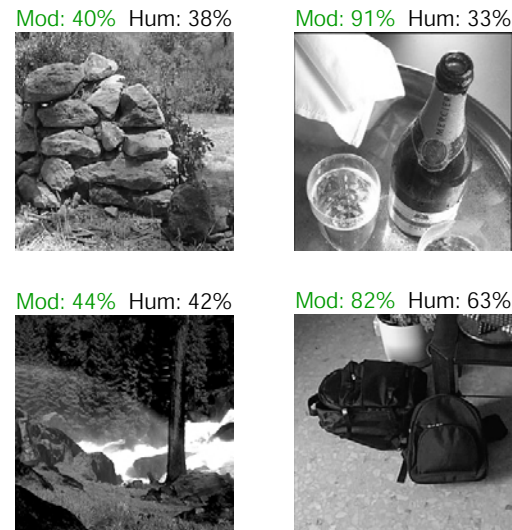


Poggio, Fig. 2

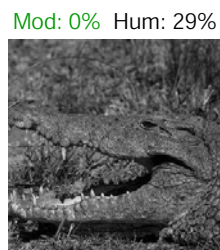
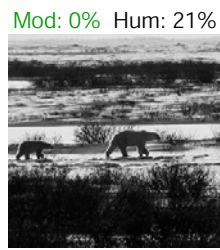
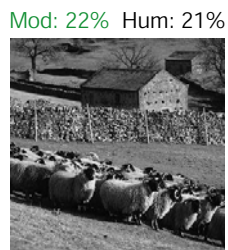




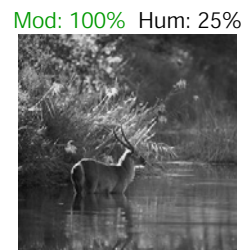
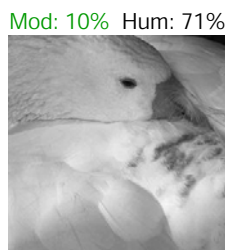
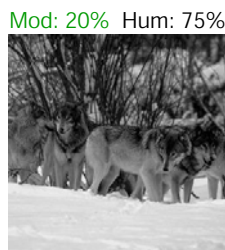
b



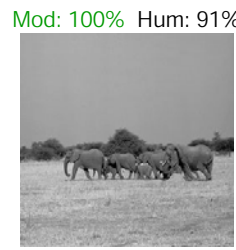
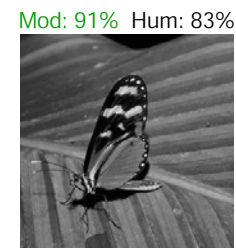
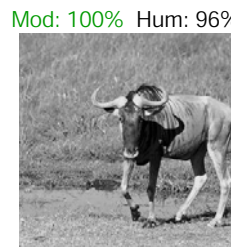
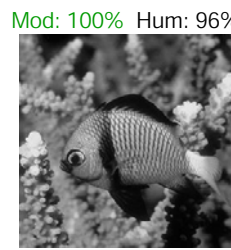
c



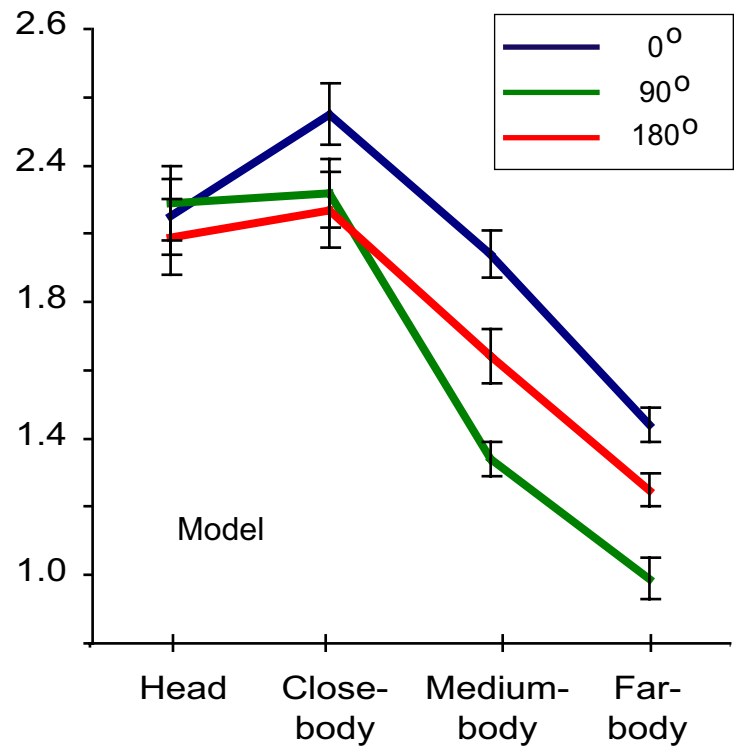
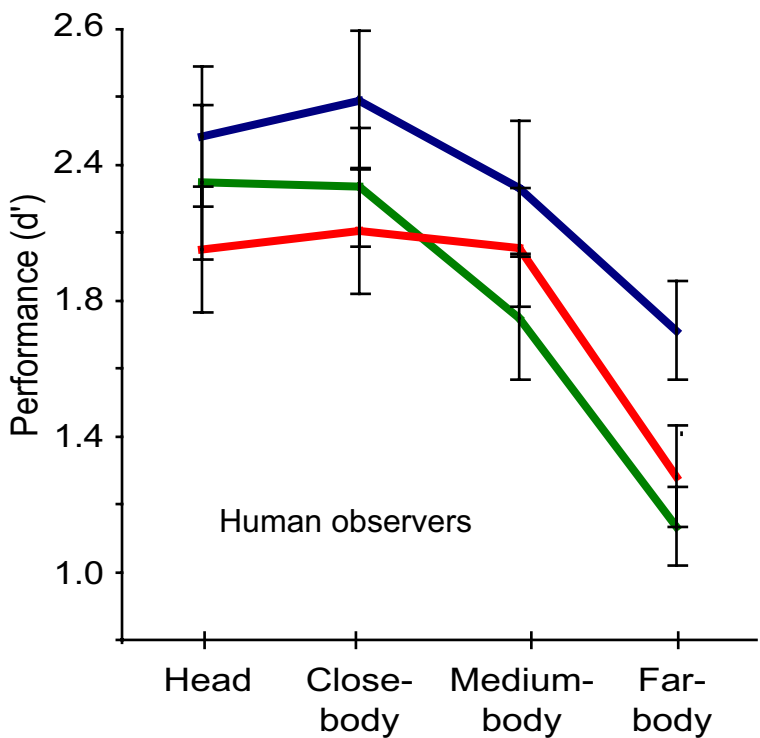
d



e



Poggio, Fig. 4



A Feedforward theory of visual cortex accounts for human performance in rapid categorization

Thomas Serre Aude Oliva Tomaso Poggio
Brain and Cognitive Sciences Department
Massachusetts Institute of Technology

Supplementary Information

Supplementary web material is also available at <http://web.mit.edu/serre/www/resources/SerreOlivaPoggio06> which includes, in particular, a basic software implementation for the model, the animal / non-animal stimulus database as well as supplementary data.

A Methods

A.1 The Model

The connectivity between model stages is described in Fig. 1 of the main manuscript. Here we briefly describe the model implementation and provide a summary of the main model parameters. A complete overview of the model can be found in [Serre et al., 2005a] and can be accessed through our Supplementary Web Material page at <http://web.mit.edu/serre/www/resources/SerreOlivaPoggio06>.

A.1.1 Architecture and Implementation

There are two types of functional layers in the model: the S layers which are composed of *simple* units are interleaved with C layers which are composed of *complex* units.

Simple units in the S_k layer pool over afferent units from a topologically related local neighborhood in the previous C_{k-1} layer with different selectivities. As a result, the complexity of the preferred stimulus of units increases from layer C_{k-1} to S_k . The pooling operation at the S level is a Gaussian-like tuning function. That is, the response y of a simple unit, receiving the pattern of synaptic inputs $(x_1, \dots, x_{n_{S_k}})$ from the previous layer is given by:

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

where σ defines the sharpness of the TUNING around the preferred stimulus of the unit corresponding to the weight vector $\mathbf{w} = (w_1, \dots, w_{n_{S_k}})$. That is, the response of the unit is maximal ($y = 1$) when the current pattern of input \mathbf{x} matches exactly the synaptic weight vector \mathbf{w} and decreases with a bell-shaped tuning profile as the pattern of input becomes more dissimilar.¹

¹When Eq. 1 is approximated by a normalized dot-product followed by a sigmoid, i.e., $y = \frac{\sum_{j=1}^{n_{S_k}} w_j x_j^p}{k + (\sum_{j=1}^{n_{S_k}} x_j^p)^r}$, the weight vector \mathbf{w} corresponds to the strength of the synaptic inputs to the Gaussian-tuned unit [see Serre et al., 2005a, pp. 11-13].

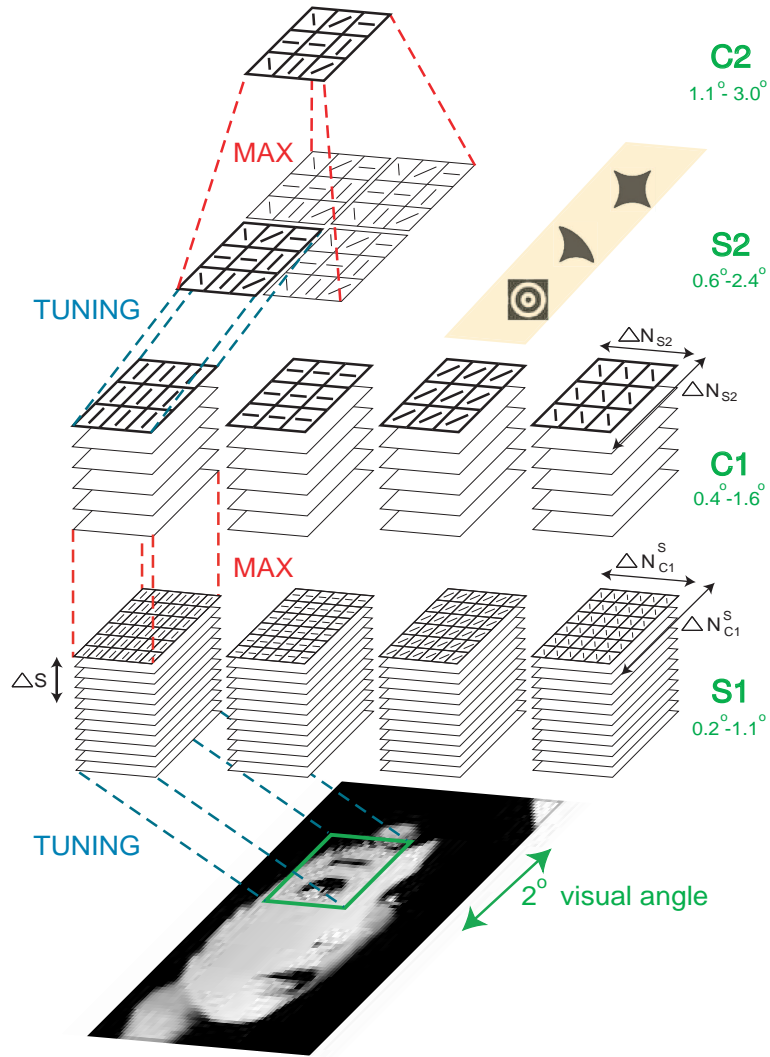


Figure S 1: A close up view in the model from S_1 to C_2 stages. The input image (gray-value) is first analyzed by an array of functionally organized S_1 units at all locations and several scales. At the next C_1 stage, a local MAX pooling operation is taken over retinotopically organized S_1 units with the *same* preferred orientation and at neighboring positions and scales to increase invariance to 2D transformations. S_2 units then combine the response of several C_1 units at *different* preferred orientations to increase the complexity of the optimal stimulus with a TUNING operation and are selective for features of moderate complexity [Kobatake et al., 1998] (examples shown in yellow). We only show one type of S_2 unit but in the model, by considering different combinations (learned from natural images) of C_1 units, we obtain $K_{S_2} \sim 2,000$ different types of S_2 units. S_2 units are also organized in feature maps such that every locations in the visual field is analyzed by all K_{S_2} types of S_2 units at different scales. A local MAX pooling operation is performed over S_2 units with the same selectivity over neighboring positions and scales to yield the C_2 unit responses. C_2 units have been shown to match well the tuning and invariance properties of cells in V4 [see Serre et al., 2005a, pp. 28-36] in response to different stimulus sets [Gallant et al., 1996; Pasupathy and Connor, 2001; Reynolds et al., 1999].

Complex units in the C_k layer pool over afferent units from the previous S_k layer with the same selectivity but at slightly different positions and scales to increase the tolerance to 2D transformations from layer S_k to C_k . The pooling operation at the complex C level is a MAX operation. That is, the response y of a complex unit corresponds to the response of the strongest of its afferents $(x_1, \dots, x_{n_{C_k}})$ from the previous S_k layer. An idealized mathematical description of the complex unit operation is given by:

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

A complete description of the two operations, a summary of the evidence as well as plausible biophysical circuits to implement them can be found in [Serre et al., 2005a, Section 5, pp. 53-59].

Functional organization: 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 Fig. S 1). Within one feature map all units share the same selectivity, *i.e.*, synaptic weight vector w which is learned from natural images (see subsection A.1.2).

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.

Simple units pool over afferent units at the same scale, *i.e.*, ΔS_{S_k} contains only a single scale element. Also note that in the current model implementation, while complex units pool over all possible afferents such that each unit in layer C_k receives $n_{C_k} = \Delta N_{C_k}^S \times \Delta N_{C_k}^S \times \Delta S_{C_k}$, 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 Table S 1 for parameter values).

Finally, 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. Note that there is a high degree of overlap between units in all stages (to guarantee good invariance to translation). The number of feature maps is conserved from S_k to C_k stage, *i.e.*, $K_{S_k} = K_{C_k}$. The value of all parameters is summarized in Table S 1.

S_1 and C_1 stages: The input to the model is a still² gray-value image ($256 \times 256 \sim 7^\circ \times 7^\circ$ of visual angle) which is first analyzed by a multi-dimensional array of simple S_1 units which correspond to the classical V1 simple cells of Hubel & Wiesel. The population of S_1 units consists in 96 types of units, *i.e.*, 2 phases \times 4 orientations \times 17 sizes (or equivalently peak spatial frequencies). Fig. S 2 shows the different weight vectors corresponding to the different types of S_1 units (only one phase shown). Mathematically the weight vector w of the S_1 units take the form of a Gabor function [Gabor, 1946], which have been shown to provide a good model of simple cell receptive fields [Jones and Palmer, 1987] and can be described by the following equation:

²The present version of the model deals with one single image at a time as it does not incorporate mechanisms for motion and the recognition of sequences. A natural extension to include time may start with a version of the original HMAX model that had the capability of recognizing image sequences [Giese and Poggio, 2003].

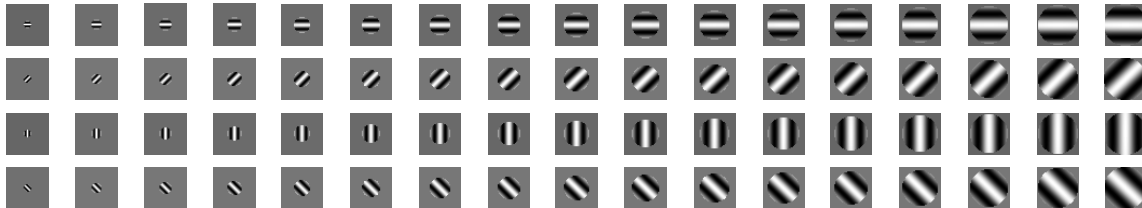


Figure S 2: The population of S_1 units: 2 phases \times 4 orientations \times 17 sizes (or equivalently peak frequencies). Only units at one phase are shown but the population also includes filters of the opposite phase. Receptive field sizes range between $0.2^\circ - 1.1^\circ$ (typical values for cortex range between $(\approx 0.1^\circ - 1^\circ$, see [Schiller et al., 1976; Hubel and Wiesel, 1965]). Peak frequencies are in the range 1.6 – 9.8 cycles/deg.

$$F(u_1, u_2) = \exp\left(-\frac{(\hat{u}_1^2 + \gamma^2 \hat{u}_2^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} \hat{u}_1\right), \quad \text{s.t.} \quad (3)$$

$$\hat{u}_1 = u_1 \cos \theta + u_2 \sin \theta \quad \text{and} \quad (4)$$

$$\hat{u}_2 = -u_1 \sin \theta + u_2 \cos \theta, \quad (5)$$

The next C_1 level corresponds to striate complex cells [Hubel and Wiesel, 1959]. 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 [Hubel and Wiesel, 1959]. As a result, the size of the receptive fields increase from the S_1 to the C_1 stage (from $0.2^\circ - 1.0^\circ$ to $0.4^\circ - 2.0^\circ$). 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 [Hubel and Wiesel, 1968; Schiller et al., 1976; DeValois et al., 1982].

The parameters of the Gabor filters (see Eq. 3) were adjusted so that the tuning properties of the corresponding S_1 units match closely those of V1 parafoveal simple cells [Serre et al., 2004]. Similarly the pooling parameters at the next stage were adjusted so that the tuning and invariance properties of the corresponding C_1 units match closely those of V1 parafoveal complex cells.³ The complete parameter set used to generate the population of S_1 units is given in Table S 1.

S_2 and C_2 stages: At the S_2 level, units pool the activities of $n_{S_2} = 10$ retinotopically organized complex C_1 units at different preferred orientations over a $\Delta N_{S_2} \times \Delta N_{S_2} = 3 \times 3$ neighborhood of C_1 units via a TUNING operation. 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 $0.6^\circ - 2.4^\circ$.

³Unlike in [Riesenhuber and Poggio, 1999], all the V1 parameters here are derived exclusively from available V1 data and do not depend as they did in part in [Riesenhuber and Poggio, 1999] from the requirement of fitting the benchmark paperclip recognition experiments [Logothetis et al., 1995]. Thus the fitting of these paperclip data by the model is even more remarkable than in [Riesenhuber and Poggio, 1999].

RF size (pixels) σ λ θ num. S_1 -types K_{S_1}	<p style="text-align: center;">S_1 parameters</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center;">7 & 9</td> <td style="text-align: center;">11 & 13</td> <td style="text-align: center;">15 & 17</td> <td style="text-align: center;">19 & 21</td> <td style="text-align: center;">23 & 25</td> <td style="text-align: center;">27 & 29</td> <td style="text-align: center;">31 & 33</td> <td style="text-align: center;">35 & 37 & 39</td> </tr> <tr> <td style="text-align: center;">2.8 & 3.6</td> <td style="text-align: center;">4.5 & 5.4</td> <td style="text-align: center;">6.3 & 7.3</td> <td style="text-align: center;">8.2 & 9.2</td> <td style="text-align: center;">10.2 & 11.3</td> <td style="text-align: center;">12.3 & 13.4</td> <td style="text-align: center;">14.6 & 15.8</td> <td style="text-align: center;">17.0 & 18.2 & 19.5</td> </tr> <tr> <td style="text-align: center;">3.5 & 4.6</td> <td style="text-align: center;">5.6 & 6.8</td> <td style="text-align: center;">7.9 & 9.1</td> <td style="text-align: center;">10.3 & 11.5</td> <td style="text-align: center;">12.7 & 14.1</td> <td style="text-align: center;">15.4 & 16.8</td> <td style="text-align: center;">18.2 & 19.7</td> <td style="text-align: center;">21.2 & 22.8 & 24.4</td> </tr> </table> <p style="text-align: center;">$0^0; 45^0; 90^0; 180^0$ 4</p>	7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37 & 39	2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2 & 19.5	3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8 & 24.4
7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37 & 39																		
2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2 & 19.5																		
3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8 & 24.4																		
Bands ΔS_{C_1} grid size ΔN_{C_1} sampling ϵ_{C_1} num. C_1 -types K_{C_1}	<p style="text-align: center;">C_1 parameters</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center;">1</td> <td style="text-align: center;">2</td> <td style="text-align: center;">3</td> <td style="text-align: center;">4</td> <td style="text-align: center;">5</td> <td style="text-align: center;">6</td> <td style="text-align: center;">7</td> <td style="text-align: center;">8</td> </tr> <tr> <td style="text-align: center;">8</td> <td style="text-align: center;">10</td> <td style="text-align: center;">12</td> <td style="text-align: center;">14</td> <td style="text-align: center;">16</td> <td style="text-align: center;">18</td> <td style="text-align: center;">20</td> <td style="text-align: center;">22</td> </tr> <tr> <td style="text-align: center;">3</td> <td style="text-align: center;">5</td> <td style="text-align: center;">7</td> <td style="text-align: center;">8</td> <td style="text-align: center;">10</td> <td style="text-align: center;">12</td> <td style="text-align: center;">13</td> <td style="text-align: center;">15</td> </tr> </table> <p style="text-align: center;">$= K_{S_1} = 4$</p>	1	2	3	4	5	6	7	8	8	10	12	14	16	18	20	22	3	5	7	8	10	12	13	15
1	2	3	4	5	6	7	8																		
8	10	12	14	16	18	20	22																		
3	5	7	8	10	12	13	15																		
grid size ΔN_{S_2} num. afferents n_{S_2} num. S_2 -types K_{S_2}	<p style="text-align: center;">S_2 parameters</p> <p style="text-align: center;">3×3 ($\times 4$ orientations)</p> <p style="text-align: center;">10 ≈ 2000</p>																								
Bands ΔS_{C_2} grid size ΔN_{C_2} sampling ϵ_{C_2} num. C_2 -types K_{C_2}	<p style="text-align: center;">C_2 parameters</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center;">1 & 2</td> <td style="text-align: center;">3 & 4</td> <td style="text-align: center;">5 & 6</td> <td style="text-align: center;">7 & 8</td> </tr> <tr> <td style="text-align: center;">8</td> <td style="text-align: center;">12</td> <td style="text-align: center;">16</td> <td style="text-align: center;">20</td> </tr> <tr> <td style="text-align: center;">3</td> <td style="text-align: center;">7</td> <td style="text-align: center;">10</td> <td style="text-align: center;">13</td> </tr> </table> <p style="text-align: center;">$= K_{S_2} \approx 2000$</p>	1 & 2	3 & 4	5 & 6	7 & 8	8	12	16	20	3	7	10	13												
1 & 2	3 & 4	5 & 6	7 & 8																						
8	12	16	20																						
3	7	10	13																						
grid size ΔN_{S_3} num. afferents n_{S_3} num. S_3 -types K_{S_3}	<p style="text-align: center;">S_3 parameters</p> <p style="text-align: center;">3×3 ($\times K_{S_2}$)</p> <p style="text-align: center;">100 ≈ 2000</p>																								
Bands ΔS_{C_3} grid size ΔN_{C_3} num. C_3 -types K_{C_3}	<p style="text-align: center;">C_3 parameters</p> <p style="text-align: center;">1 & 2 & 3 & 4 & 5 & 6 & 7 & 8</p> <p style="text-align: center;">40 $= K_{S_3} \approx 2000$</p>																								
grid size $\Delta N_{S_{2b}}$ num. afferents $n_{S_{2b}}$ num. S_{2b} -types $K_{S_{2b}}$	<p style="text-align: center;">S_{2b} parameters</p> <p style="text-align: center;">$6 \times 6; 9 \times 9; 12 \times 12; 15 \times 15$ ($\times 4$ orientations)</p> <p style="text-align: center;">100 ≈ 500 for each size ≈ 2000 total</p>																								
Bands $\Delta S_{C_{2b}}$ grid size $\Delta N_{C_{2b}}$ num. C_{2b} -types $K_{C_{2b}}$	<p style="text-align: center;">C_{2b} parameters</p> <p style="text-align: center;">1 & 2 & 3 & 4 & 5 & 6 & 7 & 8</p> <p style="text-align: center;">40 $= K_{S_{2b}} \approx 500$ for each size ≈ 2000 total</p>																								

Table S 1: Summary of all model parameters (see accompanying text).

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 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 field. Receptive field sizes at the C_2 level range between $1.1^\circ - 3.0^\circ$.

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 [Gallant et al., 1996; Pasupathy and Connor, 2001; Reynolds et al., 1999], [see Serre et al., 2005a, pp. 28-36].

S_3 and C_3 stages: 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), where the response of $n_{S_3} = 100$ C_2 units with different selectivities are combined with a TUNING operation to yield even more complex selectivities. 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 S_3 and C_3 layers provide a representation based on broadly tuned shape components.

The pooling parameters of the C_3 units (see Table S 1) 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 [Logothetis et al., 1995] (see [Serre et al., 2004, 2005a]). The receptive field sizes of the S_3 units are about $1.2^\circ - 3.2^\circ$ while the receptive field sizes of the C_3 and S_4 units is about the size of the stimulus ($7^\circ \times 7^\circ$ in the present simulation).

S_{2b} and C_{2b} stages: They may correspond to the bypass routes that have been found in visual cortex, *e.g.*, direct projections from V2 to TEO [Boussaoud et al., 1990; Nakamura et al., 1993; Gattass et al., 1997] (bypassing V4) and from V4 to TE (bypassing TEO) [Desimone et al., 1980; Saleem et al., 1992; Nakamura et al., 1993]. 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 [Hung et al., 2005] (see [Serre et al., 2005a]).

Biophysical implementations of the key computations: The model implementation used here is agnostic about the implementations of the Gaussian-like tuning and the max-like operations as well as about the biophysical mechanisms of unsupervised and supervised learning. For the two key computations we used the idealized operations described in Eq. 2 and Eq. 1. There are plausible local circuits [Serre et al., 2005a] implementing the two key operations within the time constraints of the experimental data [Perrett et al., 1992; Hung et al., 2005] based on small local population of spiking neurons firing probabilistically in proportion to the underlying analog value [Smith and Lewicki, 2006] and on shunting inhibition [Grossberg, 1973]. Other possibilities may involve spike timing in individual neurons (see [VanRullen et al., 2005] for a recent review).

A.1.2 Learning the Tuning of S Units from Natural Images

This learning stage determines 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). During this learning stage, the model becomes adapted to the statistics of the natural environment [Attneave, 1954; Barlow, 1961; Atick, 1992; Ruderman, 1994; Simoncelli and Olshausen, 2001] and units become tuned to common *image-features*⁴ that occur with high probability in natural images.

Learning in the model is sequential, *i.e.*, 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 ($\mathbf{w}^1, \dots, \mathbf{w}^{K_{S_k}}$) 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 that⁵:

$$\mathbf{w}^i = \mathbf{x} \tag{6}$$

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 positions⁶. 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*).

A.1.3 Building Task-specific Circuits from IT to PFC

We assume that a particular *program* or *routine* is set up somewhere beyond IT (possibly in PFC [Scalaidhe et al., 1999; Freedman et al., 2002, 2003; Hung et al., 2005] but the exact locus may depend on the task). Here we think of this routine as selecting a particular PFC *classification* unit (possibly a group of neurons), for a specific task. This *classification* unit combines the activity of a few hundred S_4 units (150 in the present simulation) 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 *classification* units are trained in a *supervised way*.

⁴The resulting hierarchy of unit selectivities in the model is related to other approaches such as component-based [Mohan et al., 2001; Heisele et al., 2002], part-based [Weber et al., 2000; Fergus et al., 2003; Fei-Fei et al., 2004] or fragment-based approaches [Ullman et al., 2002; Epshtein and Ullman, 2005] in computer vision. This is also sometime referred to as “bags of features” [Dorko and Schmid, 2003] in computer vision or “unbound features” [Treisman and Gelade, 1980; Evans and Treisman, 2005; Wolfe and Bennett, 1997] in cognitive science.

⁵A biophysical implementation of this rule would involve mechanisms such as LTP [Markram et al., 1997; Bi and Poo, 1998; Abarbanel et al., 2002; van Rossum et al., 2000].

⁶In 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 [Földiák, 1991].

S_4 view-tuned units: Consistent with a large body of data that suggests that the selectivity of neurons in IT depends on visual experience [Logothetis et al., 1995; Kobatake et al., 1998; Booth and Rolls, 1998; Sigala and Logothetis, 2002; Baker et al., 2002], during task-specific training of the model, S_4 units are imprinted with specific examples of the training set (25% of the training examples – both animals and non-animal images selected at random). Just like units in lower stages become tuned to patches of natural images, S_4 units become tuned to views of the target object by storing in their synaptic weights the pattern of activity of their afferents during a presentation of a particular exemplar.

PFC classification unit: In the model, the response y of a *classification* unit with input weights $\mathbf{c} = (c_1, \dots, c_{K_{S_4}})$, when presented with an input pattern $\mathbf{x} = (x_1, \dots, x_{K_{S_4}})$ from the previous layer⁷, is given by:

$$y = \sum_j c_j x_j. \quad (7)$$

The unit response $y \in \mathcal{R}$ is further binarized ($y \leq 0$) to obtain a classification label $\{-1, 1\}$. Supervised learning at this stage involves adjusting the synaptic weights \mathbf{c} so as to minimize the overall classification error on the training set E ⁸. Any linear classifier could be used at the level of the PFC units (e.g., linear SVM⁹). In this paper, we used one of the simplest types of linear classifier by computing the least-square fit solution of the regularized classification error evaluated on the training set:

$$E = \sum_{i=1}^l \|y^i - \hat{y}^i\|^2 + \lambda \|\mathbf{c}\|^2. \quad (8)$$

where y^i corresponds to the classification unit response for the i^{th} training example, \hat{y}^i is the true label of the i^{th} training example and λ is a fixed constant. To solve Eq. 8 we used the non-biological Matlab© (The MathWorks, Inc) left division operation for matrices but we obtained similar results with a more biologically plausible stochastic gradient learning approach using weight perturbations modified from [Sutton and Barto, 1981; Seung, 2003].

Remark: Both the supervised and unsupervised learning stages are relatively fast. Yet at runtime, it takes about one minute to classify a single image. A speed up by a factor of 10 is feasible.

A.2 Animal vs. Non-Animal Categorization Task

A.2.1 Definition of d' Sensitivity Measure

d' is a standard sensitivity measure of performance [Macmillan and Creelman, 1991; Sekuler and Blake, 2002] which, for each observer, combines both the *hit* rate H , i.e., the proportion of animal images correctly classified by the observer, and the *false alarm* rate F , i.e., the proportion of non-animal images incorrectly classified by the observer, into one single standardized score. The mathematical form of the d' measure is:

⁷Recall that S_4 unit j , denoted x_j , is tuned to the j^{th} training example.

⁸While only 25% of the training set is stored at the S_4 level, the full training set is used to adjust the synaptic weights of the classification unit.

⁹A recent study [Hung et al., 2005] demonstrated that a linear classifier can indeed *read-out* with high accuracy and over extremely short times (a single bin as short as 12.5 millisecond) object identity, object category and other information (such as position and size of the object) from the activity of about 100 neurons in IT.

$$d' = Z(H) - Z(F), \quad (9)$$

where Z corresponds to the inverse of the normal distribution function, *i.e.*, Z transforms a hit or false alarm rate to a z -score, that is, to standard-deviation units.

The d' measure is a monotonic function of the performance of the observer. For instance, a good observer has a low F rate and a high H rate and therefore the d' is large. A poor observer tends to have a higher F rate and/or a lower H rate and the d' is smaller. A system for which $F = H$ is at chance and has $d' = 0$.

A.2.2 Random Split Procedures

They are common practice in machine learning for evaluating the performance of a classifier as they have been shown to provide a good estimate of the expected error of a classifier [Devroye et al., 1996]. The procedure is as follow:

1. Split the set of 1,200 (animal and non-animal) images into two halves. Denote one set *Training* and the other *Test*.
2. Train the classifier on the labeled *Training* set of images.
3. Evaluate the performance of the classifier on the *Test* set of images.

The procedure above was repeated N times and performance was averaged across these N random runs.

A.2.3 The Stimulus Dataset

All images were gray-value 256×256 pixel images. The stimulus database contains a total of 600 animal stimuli and 600 non-animal stimuli. The 600 animal stimuli were extracted from the Corel database as in [Thorpe et al., 1996]. 256×256 image windows were cropped around the animal from the original 256×384 pixel images with a random offset to prevent the animal from always be presented in the center of the image. Animal images were manually grouped into four categories with 150 exemplars in each, *i.e.*, *head*, *close-body*, *medium-body* and *far-body*.

A set of distractors with matching mean distance from the camera (300 from natural and 300 from artificial scenes) was selected from a database of annotated mean depth images [Torralba and Oliva, 2002]. We selected images with a mean distance from the camera below 1 m for head, between 5 m and 20 m for close-body, between 50 m and 100 m for medium-body as well as above 100 m and panoramic views for far-body. The database is publicly available at <http://web.mit.edu/serre/www/resources/SerreOlivaPoggio06>.

A.2.4 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 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 [Renninger and Malik, 2004]

	Head	Close-body	Medium-body	Far-body
Mean luminance	0.28	0.36	0.46	0.34
Gray value SVM	0.23	0.22	0.17	0.13
Textons	0.84	0.58	0.69	0.35
Global features	1.43	1.73	1.47	0.74
Model C_1 layer	1.37	1.78	1.53	0.65

Table S 2: Summary of the performance of several benchmark computer vision systems on the animal database. The poor performance of simple classification strategies indicate that it is very unlikely that human observers could rely on low-level cues. The performance of the model is significantly higher than all of the benchmarks with a d' of 2.04, 2.48, 1.97, 1.37 on the subcategories.

and a *global feature*-based system [Torralba and Oliva, 2003]. 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 at <http://web.mit.edu/serre/www/resources/SerreOlivaPoggio06>.

The performance of the different approaches is summarized in Table S 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 [Renninger and Malik, 2004; Torralba and Oliva, 2003] as well as the model C_1 layer perform significantly better, their level of performance remains significantly 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 Serre et al., 2005a, 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 Fig. S6 for a comparison between different layers of the model on the animal / on-animal categorization task).

A.2.5 Categorization by the Human Observers

For all three experiments, participants gave a written informed consent. All participants were between 18 and 35 years old, with $n = 24, 14$ and 21, in experiments 1, 2 (see main text) and 3 (see section D.1) respectively. There was approximately the same number of male and female observers in each experiment and none participated in more than one of the three experiments. Participants were seated in a dark room, 0.5 m away from a computer screen, connected to a computer (Intel Pentium®IV processor, 1 GB RAM, 2.4 GHz). The monitor refresh rate was 100 Hz allowing stimuli to be displayed with a frame-duration of 10 ms and a resolution of 1024×768 .

We used the Matlab® (Mathworks Inc, Natick, MA) software with the psychophysics toolbox [Brainard, 1997; Pelli, 1997] to precisely time the stimulus presentations. In all experiments, the image duration was 20 ms. In experiments 1 and 2, the mask appeared after a fixed inter-stimulus interval (ISI) of 30 ms (corresponding to a Stimulus Onset Asynchrony SOA of 50 ms). In experiment 3, we tested variable $ISIs$ of 0 ms (thus corresponding to an SOA of 50 ms). In experiment 3, we randomly interleaved different ISI conditions: 0 ms ISI ($SOA = 20$ ms), 30 ms ISI ($SOA = 50$ ms), 60 ms ISI ($SOA = 80$ ms), or infinite (*i.e.*, never appeared). The mask following

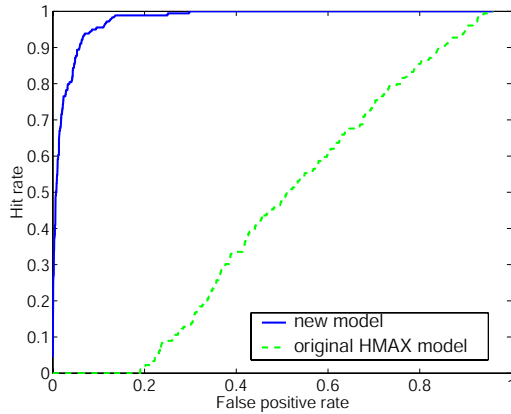


Figure S 3: Face detection in natural images: Comparison between the “original HMAX model” [Riesenhuber and Poggio, 1999] and the new model (reproduced from [Serre et al., 2002]).

the picture was a $(1/f)$ random noise mask, generated (for each trial) by filtering random noise through a Gaussian filter.

The stimuli were presented in the center of the screen (256×256 pixels, about $7^\circ \times 7^\circ$ of visual angle, gray-level images). The 1,200 image stimuli (600 animals and 600 distractors) were presented in random order and divided into 10 blocks of 120 images each. Participants were asked to answer as fast and as accurately as possible if the image contained an animal, by pressing a *yes* or *no* key on a computer keyboard. They were randomly asked to use their left or right hand for *yes* vs. *no* answers. Each experiment took about thirty minutes to perform.

A.2.6 Categorization by the Model

To train the PFC classification unit in the model, we used a random splits procedure (see subsection A.2.2) with $n = 20$ random runs. As described in subsection A.1.3 we trained the classifier on a set of training examples as (x^i, y^i) pairs, where x^i denotes the i^{th} image in the training set and y^i its associated label (animal or non-animal). In each run, half of the 1,200 images from the database of stimuli in experiment 1, 2 and 3 was used for training the model and the remaining half for testing it. The model performance reported in experiment 1, 2, and 3 was averaged over these n random runs. Note that the error bars for the model in Fig. 3 and 4 in the main manuscript correspond to the standard errors computed over these $n = 20$ random runs. Note that a single classifier was trained on all four animal and non-animal categories together.

B Comparison with Other Approaches

B.1 Comparison with the Original Model

In the original model [Riesenhuber and Poggio, 1999], learning only occurred in the top-most layers (*i.e.*, the task specific circuits) and units in intermediate layers were hardwired (simple 2×2 combinations of 4 orientations). The present model contains a stage of unsupervised learning from S_2 to S_4 stages, which is more faithful to the physiology data and performs significantly better in recognizing real-world images (see Fig. S 3).

B.2 Comparison with Other Computer Vision Systems

Table S 3 summarizes several comparisons between the model and other state-of-the-art computer vision systems. For this comparison, an earlier (simpler) implementation of the model [Serre et al., 2005b], which corresponds to the *bypass* route projecting from $S_1 \rightarrow C_1 \rightarrow S_{2b} \rightarrow C_{2b}$, was used. The performance of the full architecture which includes a richer dictionary of shape components, tends to be significantly higher than the performance of this simpler (incomplete) implementation. Therefore the results reported here constitute a lower bound on the system performance. These comparisons are based on three studies¹⁰:

- In [Serre et al., 2005b] we compared the model to the constellation models [Weber et al., 2000; Fergus et al., 2003] on five standard publicly available datasets from the Caltech vision group: Leave (*Lea*), Car, Face (*Fac*), Airplane (*Air*) and Motorcycle (*Mot*) as well as two other component-based systems [Heisele et al., 2002; Leung, 2004] on the MIT-CBCL Face (*Fac*) and Car datasets.
- [Chikkerur and Wolf, 2006] re-implemented the fragment-based system by Ullman and colleagues [Ullman et al., 2002; Epshtein and Ullman, 2005] for comparison with the model on five publicly available datasets: the Leave, Face and Motorcycle datasets from CalTech and the Cow and Face dataset from the Weizmann Institute.
- [Bileschi and Wolf, 2005] re-implemented several systems for comparison with the model on the MIT-CBCL Street Scene dataset. They re-implemented two object recognition systems [Torralba et al., 2004; Leibe et al., 2004] for comparison on the “shape-based” object categories, *i.e.*, Bike (*Bik*), Pedestrian (*Ped*), and Car as well as two texture recognition systems [Renninger and Malik, 2004; Carson et al., 1999] for comparison on the “texture-based” object categories, *i.e.*, Building (*Bui*), Tree (*Tre*), Road (*Roa*) and Sky.

In Table S 3, blue indicates that the corresponding study [Serre et al., 2005b] relied on published results of the benchmark systems on standard datasets. Yellow indicates that the results for the benchmark systems were based on re-implementations by the authors of the studies [Bileschi and Wolf, 2005; Chikkerur and Wolf, 2006]. In the study by [Bileschi and Wolf, 2005] the two numbers for the model on Bike, Pedestrian and Car correspond to the performance of the model C_{2b} and C_1 units respectively.¹¹

The system by Ullman and colleagues [Ullman et al., 2002; Epshtein and Ullman, 2005] along with our own model are the only biologically plausible approaches that have been extensively tested on standard computer vision datasets. There exist older feedforward models of object recognition in cortex that have been designed not as high performance systems but rather as insightful proofs of concept (*e.g.*, [Wallis and Rolls, 1997; Mel, 1997]) and were not tested on complex recognition tasks. In fact it is shown in [Mel, 1997] that the simple addition of a randomly colored lowercase character in the target image (acting as clutter) is enough to drive a seven-fold increase in the error rate of the system. As discussed by the author (pp 801), one way to overcome these limitations is by “by conjoining existing (or other) low-order features into compound features of higher order. The issue of feature locality remains crucial, however.”, which is precisely what our current model does.

¹⁰Mutch & Lowe also reports favorable comparison in their implementation of the model [Mutch and Lowe, 2006].

¹¹On these datasets, images are aligned and normalized, and the amount of clutter is minimal. For such tasks, for which there is no variation of the object in shift and scale, lower stages of the model (*e.g.*, C_1 stage) tend to perform better than higher stages (*e.g.*, C_{2b}).

		Weizmann		CalTech			MIT-CBCL			
		Fac	Cow	Lea	Car	Fac	Air	Mot	Fac	Car
[Serre et al, 2005]	Model [Serre et al, 2005]			97.0	99.7	98.2	96.7	98.0	95.9	95.1
	Constellation [Weber et al, 2000, Fergus et al, 2003]			84.0	84.8	96.4	94.0	95.0		
	Component-based [Heisele et al, 2002]								90.4	
	Component-based [Leung, 2004]									75.4
[Chikkerur & Wolf, 2006]	Model [Serre et al, 2005]	100.0	92.0	97.9		94.5		96.5		
	Fragments [Epshtein & Ullman, 2005]	98.0	78.7	87.4		66.8		52.6		
	Single template SVM	100.0	77.3	71.6		62.2		65.6		
		MIT-CBCL Street Scene Database								
		Bik	Ped	Car		Bui	Tre	Roa		Sky
[Bileschi & Wolf, 2005]	Model [Serre et al, 2005]	87.8	81.7	89.6		80.3	90.8	88.9		94.7
		84.1	88.8	92.9						
	Component-based [Torralba et al, 2004]	68.5	79.8	69.9						
	Part-based [Leibe et al, 2004]	80.9	85.2	85.9						
	Single template SVM	67.8	70.0	85.0						
	Blobworld [Carson et al, 1999]					66.1	85.8	73.1		68.2
	Texton [Renninger & Malik, 2002]					69.7	70.4	58.1		65.1
	Histogram of edges					63.3	63.7	73.3		68.3

Table S 3: Summary of the comparisons between the model and other computer vision systems (see text).

C Supplementary Discussions

C.1 On the Number of Units in the Model

The model described in Fig. 1 contains on the order of 10 million units. Given the biophysical circuits described in [Serre et al., 2005a], one unit in the model may correspond to a group of $\sim 10 - 100$ neurons in cortex. This estimate results in about $10^8 - 10^9$ actual neurons recruited by the model, which corresponds to about 0.01% to 1% of visual cortex (based on 10^{11} neurons in cortex [Kandel et al., 2000]). While the number of units in the model is very large it remains much smaller than the proportion of cortex taken by visual areas.¹²

Obviously our model does not constitute a complete model of visual cortex. In particular the model only accounts for the ventral stream and several other processing channels such as color, motion and stereo would have to be incorporated. Yet taken all of these limitations into account and assuming that we need to increase our estimate of the number of neurons by one order of magnitude, it remains that the model can categorize visual object with no more than 10% of visual cortex.

C.2 On Interrupting Back-Projections with the Mask

There is much debate about the effect of a mask – as used in the psychophysics described here – on visual processing. A well accepted theory is the “interruption theory” that has been in fact corroborated by physiological studies [Rolls and Tovee, 1994; Tovee, 1994; Kovács et al., 1995; Rolls et al., 1999; Keysers et al., 2001] (see also [Lamme and Roelfsema, 2000]). 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).

Here we would like to try to isolate a purely feedforward sweep from further recurrent processing [Lamme and Roelfsema, 2000]. Whether or not the back-projections may participate in the overall processing and contribute to the final performance 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* to stage *B* and back to stage *A* is longer than the *SOA*, this back-projection will not influence the processing in the visual system as it will be interrupted before.

Based on estimates of conduction delays (see Fig. S 4), extrapolated from monkey [Nowak and Bullier, 1997; Thorpe and Fabre-Thorpe, 2001] to human [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 active¹³, for instance, between IT and V4 (see Fig. S 4, 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 level¹⁴.

¹²If we were to train the model to recognize a more plausible number of discriminable objects (*i.e.*, $\sim 30,000$ [Biederman, 1987]), the number of units in the model would increase by another 10 million. This comes from the fact that the same basic dictionary of shape-tuned units (*i.e.*, from S_1 up to S_4) is being used for different recognition tasks resulting in only a small number of extra units being added in the top layers for each new category.

¹³Note that for such *SOA*, local feedback loops green arrows in Fig. S 4 are likely to be already active ($\Delta < 20 - 30$ *ms*), see [Knierim and van Essen, 1992; Zhou et al., 2000].

¹⁴The mask is likely to interrupt the maintained response of IT neurons but not to alter their initial selective response [Kovács et al., 1995; Rolls et al., 1999]. According to an independent study [Hung et al., 2005] this would provide significantly more time than needed ($\gg 12.5$ *ms*) to permit robust recognition in “reading out” from monkey IT neurons.

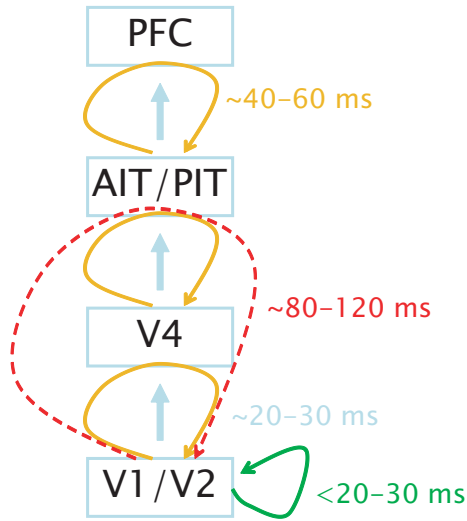


Figure S 4: Estimate of the timing of feedback loops in the ventral stream of primates (based on [Nowak and Bullier, 1997; Thorpe and Fabre-Thorpe, 2001]). We assume that typical latencies from one stage to the next is $\sim 10\text{ ms}$ and that feedforward and back-projections have similar conduction times [Nowak and Bullier, 1997]. The first number corresponds to latencies for monkeys and is assumed to constitute a lower bound on the latencies for humans. The second number corresponds to an additional 50% and is assumed to constitute a “typical” number for humans [Thorpe, Personal communication].

This estimate seems in good agreement with results from a Transcranial Magnetic Stimulation (TMS) experiment [Corthout et al., 1999] that has shown a disruption of the feedforward sweep [Lamme and Roelfsema, 2000] for pulses applied between 30 ms and 50 ms after stimulus onset.¹⁵ It is thus quite interesting that the model matches human performance almost exactly for an SOA of 50 ms , but underperforms it for longer $SOAs$ (see subsection D.1 below). 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 backprojections in visual processing can be found in [Lamme and Roelfsema, 2000; Hochstein and Ahissar, 2002].

D Supplementary Data

D.1 Experiment 3: Varying the SOA

We replicated previous psychophysical results [Bacon-Mace et al., 2005] to test the influence of the mask on visual processing with four experimental conditions, *i.e.*, when the mask followed the target image (20 ms presentation): a) without any delay (“immediate-mask” condition); b) with a short inter-stimulus interval of 30 ms (50 ms SOA); c) with an ISI of 60 ms (80 ms SOA) or d) never (“no-mask” condition). A comparison between the performance of human observers ($n = 21$) and the model is shown in Fig. S 5. Here we plot the hit rate as error measure: 1) for comparison with previous studies with go/no-go tasks [Thorpe et al., 1996; VanRullen and Koch, 2003; Bacon-Mace et al., 2005] that report the hit rate as error measures and 2) to emphasize the fact that false alarm rates of the human observers is about constant across the various backward masking conditions (16%, 16%, 16% and 14%) and the improvement in d' (see <http://web.mit.edu/serre/www/resources/SerreOlivaPoggio06>) comes from an increase in the hit rates for human observers.

¹⁵The same experiment [Corthout et al., 1999] also demonstrated blockade of perception by pulses applied between $80 - 120\text{ ms}$, presumably corresponding to recurrent processing [Lamme and Roelfsema, 2000] by the back-projections.

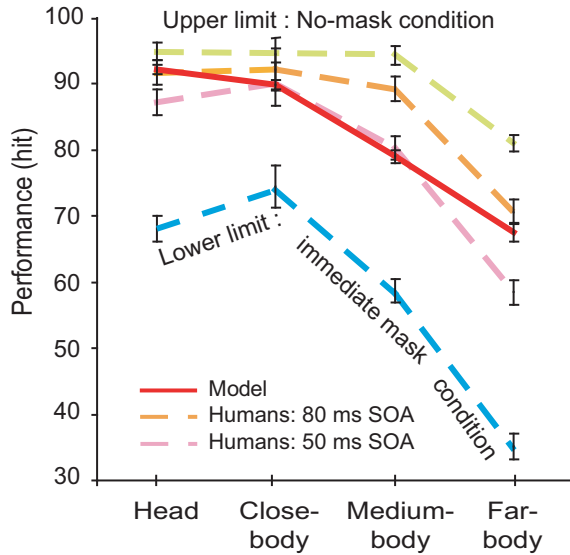


Figure S 5: Comparison between the model and human observers with different mask conditions: The upper and lower bounds on human-level performance ($n = 21$) are given by the no-mask condition and the immediate-mask condition respectively. The average accuracy of human observers for the conditions with immediate-mask, 50 ms SOA, 80 ms SOA and no-mask conditions were 59%, 79%, 86% and 91% respectively - all significantly above chance (t-test, $p < 0.01$) - compared to 82% for the model (18% false-alarms). The model matches human observers for SOAs between 50 ms and 80 ms. Error bars indicate the standard error and are not directly comparable for the model (computed over $n = 20$ random runs) and for humans (computed over $n = 21$ observers).

We found that the accuracy of the human observers was well within the range of data previously obtained with go/no-go tasks [Thorpe et al., 1996; VanRullen and Koch, 2003; Bacon-Mace et al., 2005]. The level of performance of human observers reached a ceiling in the 80 ms SOA condition (except when the animal was camouflaged in the scene, i.e. far-body group). As expected, the SOA modulates the level of performance of the observers, improving from a low level of recognition in the immediate-mask condition to ceiling in the no-mask condition. The model predicts human-level hit rate very well between the 50 ms SOA and the 80 ms SOA conditions.

D.2 A Comparison Between Layers of the Model

Fig. S6 shows the performance of different layers of the model on the animal *vs.* non-animal categorization task and the Human observers ($n=24$, SOA = 50 ms). Full model corresponds to the complete model of Fig. 1, Bypass route only corresponds to an implementation of the model for which V4 was lesioned, Direct route only corresponds to an implementation of the model for which the route from V2 to PIT (bypassing V4) was lesioned, V1 only corresponds to classifying the C_1 layer alone. The performance of all the various model implementations were obtained with $n = 10$ random splits.

D.3 Supplementary Web Material

A summary of all performance measures for both human observers and the model (including an ROC analysis, error and hit rates) as well as reaction times for human observers are available in the Supplementary Web Material at <http://web.mit.edu/serre/www/resources/SerreOlivaPoggio06>.

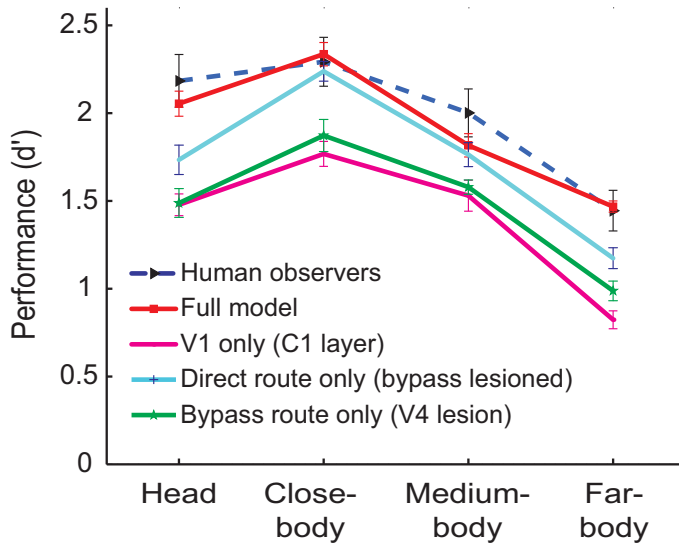


Figure S 6: Comparison between different layers of the model on the animal *vs.* non-animal categorization task. The poor performance of the model after lesioning V4 is likely to come from a lack of invariance to position and scale at the level of IT.

Supplementary References

- H.D.I. Abarbanel, R. Huerta, and M.I. Rabinovich. Dynamical model of long-term synaptic plasticity. *Proc. Nat. Acad. Sci. USA*, 99(15):10132–10137, 2002.
- J.J. Atick. Could information theory provide an ecological theory of sensory processing. *Network: Computation in Neural Systems*, 3:213–251, 1992.
- F. Attneave. F. 1954. some informational aspects of visual perception. *Psychol. Rev.*, 61:183–193, 1954.
- N. Bacon-Mace, M.J. Mace, M. Fabre-Thorpe, and S.J. Thorpe. The time course of visual processing: backward masking and natural scene categorisation. *Vis. Res.*, 45:1459–1469, 2005.
- C.I. Baker, M. Behrmann, and C.R. Olson. Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.*, 5:1210–1216, 2002.
- H.B. Barlow. *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, pages 217–234. MIT Press, Cambridge, MA, wa rosenblith edition, 1961.
- G.Q. Bi and M.M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, 18:10464–10472, 1998.
- I. Biederman. Recognition-by-components: A theory of human image understanding. *Psych. Rev.*, 94:115–147, 1987.
- S. Bileschi and L. Wolf. A unified system for object detection, texture recognition, and context analysis based on the standard model feature set. In *Proc. British Machine Vision Conference*, 2005.
- M. C. Booth and E. T. Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex*, 8:510–523, 1998.

- D. Boussaoud, L. G. Ungerleider, and R. Desimone. Pathways for motion analysis: cortical connections of the medialsuperior temporal and fundus of the superior temporal visual areas in themacaque. *J. Comp. Neurol.*, 296(3):462–95, June 1990.
- D.H. Brainard. The psychophysics toolbox. *Spat. Vis.*, 10:433–436, 1997.
- C. Carson, M. Thomas, S. Belongie, J. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*, 1999.
- S. Chikkerur and L. Wolf. Empirical comparison between hierarchical fragments based and standard model based object recognition systems. Cbcl paper mmvi-0i, MIT, 2006.
- E. Corthout, B. Uttl, V. Walsh, M. Hallett, and A. Cowey. Timing of activity in early visual cortex as revealed by transcranial magnetic stimulation. *Neuroreport*, 1999.
- R. Desimone, J. Fleming, and C.D. Gross. Prestriate afferents to inferior temporal cortex: an hrp study. *Brain Res.*, 184:41–55, 1980.
- R.L. DeValois, D.G. Albrecht, and L.G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:545–559, 1982.
- L. Devroye, G. Laszlo, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New York, 1996.
- G. Dorko and C. Schmid. Selection of scale invariant neighborhoods for object class recognition. In *International Conference on Computer Vision*, 2003.
- B. Epshtein and S. Ullman. Feature hierarchies for object classification. In *Proc. of the International Conference on Computer Vision*, pages 220–227, 2005.
- K.K Evans and A. Treisman. Perception of objects in natural scenes: Is it really attention free? *J. Exp. Psych.: Hum. Percept. Perf.*, 31(6):1476–1492, 2005.
- L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Generative-Model Based Vision*, 2004.
- R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2003.
- P. Földiák. Learning invariance from transformation sequences. *Neural Comp.*, 3:194–200, 1991.
- D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *J. Neurophys.*, 88:930–942, 2002.
- D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. Comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.*, 415:5235–5246, 2003.
- D. Gabor. Theory of communication. *J. IEE*, 93:429–459, 1946.

- J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. van Essen. Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. *J. Neurophys.*, 76: 2718–2739, 1996.
- R. Gattass, A.P. Sousa, M. Mishkin, and L.G. Ungerleider. Cortical projections of area v2 in the macaque. *Cereb. Cortex*, 7:110–129, 1997.
- M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience*, 4:179–192, 2003.
- S Grossberg. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52:213–257, 1973.
- B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Advances in Neural Information Processing Systems*, volume 14, 2002.
- S. Hochstein and M. Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36:791–804, 2002.
- D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.*, 28:229–289, 1965.
- D. H. Hubel and T. N. Wiesel. Receptive fields of single neurons in the cat’s striate visual cortex. *J. Phys.*, 148:574–591, 1959.
- D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. Phys.*, 195:215–243, 1968.
- C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo. Fast read-out of object identity from macaque inferior temporal cortex. *Science*, 310:863–866, November 2005.
- J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophys.*, 58:1233–1258, 1987.
- E. R. Kandel, J. H. Schwartz, and T. M. Jessell. *Principles of Neural Science*. McGraw-Hill Companies, Inc., 2000.
- C. Keysers, D. K. Xiao, P. Földiák, and D. I. Perrett. The speed of sight. *J. Cogn. Neurosci.*, 13:90–101, 2001.
- J.J. Knierim and D.C. van Essen. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophys.*, 67(4):961–p80, 1992.
- E. Kobatake, G. Wang, and K. Tanaka. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophys.*, 80:324–330, 1998.
- G. Kovács, R. Vogels, and G.A. Orban. Cortical correlate of pattern backward masking. *Proc. Nat. Acad. Sci. USA*, 92:5587–5591, 1995.
- V.A.F. Lamme and P.R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosci.*, 23:571–579, 2000.

- B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *SLCP '04 Workshop on Statistical Learning in Computer Vision*, 2004.
- B. Leung. Component-based car detection in street scene images. Master's thesis, EECS, MIT, 2004.
- N. K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5:552–563, 1995.
- N. A. Macmillan and C. D. Creelman. *Detection Theory: A User's Guide*. Cambridge University Press, 1991.
- H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215, 1997.
- B. W. Mel. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comp.*, 9:777–804, 1997.
- A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 23, pages 349–361, 2001.
- J. Mutch and D. Lowe. Multiclass object recognition using sparse, localized hmax features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. To appear.
- H. Nakamura, R. Gattass, R. Desimone, and L. G. Ungerleider. The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J. Neurosci.*, 13(9):3681–3691, September 1993.
- L.G. Nowak and J. Bullier. *Extrastriate visual cortex in primates*, volume 12, chapter The timing of information transfer in the visual system, pages 205–241. New York: Plenum Press, 1997.
- A. Pasupathy and C. E Connor. Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophys.*, 86(5):2505–2519, 2001.
- D.G. Pelli. The video toolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.*, 1997.
- D.I. Perrett, J.K. Hietanen, M.W. Oram, and P.J. Benson. Organization and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. Roy. Soc. B*, 335:23–30, 1992.
- L.W. Renninger and J. Malik. When is scene identification just texture recognition? *Vis. Res.*, 44: 2301–2311, 2004.
- J. H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.*, 19:1736–1753, 1999.
- M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2:1019–1025, 1999.
- E.T. Rolls and M.J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. Royal Society of London B*, 1994.

- E.T. Rolls, M.J. Tovee, and S. Panzeri. The neurophysiology of backward visual masking: information analysis. *JCN*, 11:300–311, 1999.
- D. Ruderman. The statistics of natural images. *Network : Computation in Neural Systems*, 5:598–605, 1994.
- K.S. Saleem, K. Tanaka, and K.S. Rockland. Pha-I study of connections from teo and v4 to te in the monkey visual cortex. *Society for Neuroscience Abstracts*, 18(294), 1992.
- S.P.O. Scalaidhe, F.A.W. Wilson, and P.S. Goldman-Rakic. Areal segregation of face-processing neurons in prefrontal cortex. *Science*, 278(5340):1135–1138, 1999.
- P. H. Schiller, B. L. Finlay, and S. F. Volman. Quantitative studies of single-cell properties in monkey striate cortex I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, 39(6):1288–1319, 1976.
- R. Sekuler and R. Blake. *Perception*. McGraw Hill, fourth edition, 2002.
- T. Serre, M. Riesenhuber, J. Louie, and T. Poggio. On the role of object-specific features for real world object recognition. In S.-W. Lee, H. H. Buelthoff, and T. Poggio, editors, *Proc. of Biologically Motivated Computer Vision*, Lecture Notes in Computer Science, New York, 2002. Springer.
- T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. AI Memo 2004-026 / CBCL Memo 243, MIT, Cambridge, MA, 2004.
- T. Serre, M. Kouh., C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. AI Memo 2005-036 / CBCL Memo 259, MIT, Cambridge, MA, 2005a.
- T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In IEEE Computer Society Press, editor, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, 2005b.
- H.S. Seung. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40:1063–1073, 2003.
- N. Sigala and N. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320, 2002.
- E.P. Simoncelli and B.A. Olshausen. Natural image statistics and neural representation. *Ann. Rev. Neurosci.*, 24:1193–1216, 2001.
- E.C. Smith and M.S. Lewicki. Efficient auditory coding. *Nature*, 2006.
- R.S. Sutton and A.G. Barto. Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.*, 88:135–170, 1981.
- S.J. Thorpe, Personal communication.
- S.J. Thorpe and M. Fabre-Thorpe. Seeking categories in the brain. *Science*, 291:260–263, 2001.
- S.J. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

- A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Pattern Analysis and Machine Intelligence*, 24, 2002.
- A. Torralba and A. Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14:391–412, 2003.
- A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.
- M.J. Tovee. Neuronal processing. how fast is the speed of thought? *Curr. Biol.*, 1994.
- A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cog. Psych.*, 12:97–136, 1980.
- S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.*, 5(7):682–687, 2002.
- M.C.W. van Rossum, G.Q. Bi, and G.G. Turrigiano. Stable hebbian learning from spike timing-dependent plasticity. *J. Neurosci.*, 20(23):8812–8821, 2000.
- R. VanRullen and C. Koch. Visual selective behavior can be triggered by a feed-forward process. *JCN*, 15:209–217, 2003.
- R. VanRullen, R. Guyonneau, and S.J. Thorpe. Spike times make sense. *Trends in Neurosci.*, 28(1), 2005.
- G. Wallis and E. T. Rolls. A model of invariant object recognition in the visual system. *Prog. Neurobiol.*, 51:167–194, 1997.
- M. Weber, W. Welling, and P. Perona. Unsupervised learning of models of recognition. In *Proc. of the European Conference on Computer Vision*, volume 2, pages 1001–1008, 2000.
- J.M. Wolfe and S.C. Bennett. Preattentive object files: shapeless bundles of basic features. *Vis. Res.*, 37:25–44, 1997.
- H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20:6594–6611, 2000.