A feedforward architecture accounts for rapid categorization

Thomas Serre[*,†,‡,§], Aude Oliva[‡], and Tomaso Poggio[*,†,‡]

[*]Center for Biological and Computational Learning, [†]McGovern Institute for Brain Research, and [‡]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by Richard M. Held, Massachusetts Institute of Technology, Cambridge, MA, January 26, 2007 (received for review November 11, 2006)

Primates are remarkably good at recognizing objects. The level of performance of their visual system and its robustness to image degradations still surpasses the best computer vision systems despite decades of engineering effort. In particular, the high accuracy of primates in ultra rapid object categorization and rapid serial visual presentation tasks is remarkable. Given the number of processing stages involved and typical neural latencies, such rapid visual processing is likely to be mostly feedforward. Here we show that a specific implementation of a class of feedforward theories of object recognition (that extend the Hubel and Wiesel simple-to-complex cell hierarchy and account for many anatomical and physiological constraints) can predict the level and the pattern of performance achieved by humans on a rapid masked animal vs. non-animal categorization task.

object recognition | computational model | visual cortex | natural scenes | preattentive vision

Object recognition in the cortex is mediated by the ventral visual pathway running from the primary visual cortex (V1) (1) through extrastriate visual areas II (V2) and IV (V4), to the inferotemporal cortex (IT) (2–4), and then to the prefrontal cortex (PFC), which is involved in linking perception to memory and action. Over the last decade, a number of physiological studies in nonhuman primates have established several basic facts about the cortical mechanisms of recognition. The accumulated evidence points to several key features of the ventral pathway. From V1 to IT, there is an increase in invariance to position and scale (1, 2, 4–6) 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, 7). Finally, plasticity and learning are probably present at all stages and certainly at the level of IT (6) and PFC.

However, an important aspect of the visual architecture, i.e., the role of the anatomical back projections abundantly present between almost all of the areas in the visual cortex, remains a matter of debate. The hypothesis that the basic processing of information is feedforward is supported most directly by the short time spans required for a selective response to appear in IT cells (8). Very recent data (9) show that the activity of small neuronal populations in monkey IT, over very short time intervals (as small as 12.5 ms) and only ≈100 ms after stimulus onset, contains surprisingly accurate and robust information supporting a variety of recognition tasks. Although this finding does not rule out 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 (10–13).

One of the first feedforward models, Fukushima's Neocognitron (14), followed the basic Hubel and Wiesel proposal (1) for building an increasingly complex and invariant object representation in a hierarchy of stages by progressively integrating convergent inputs from lower levels. Building on several existing neurobiological models (5, 15–19, ¶), conceptual proposals (1, 2, 20, 21), and computer vision systems (14, 22), we have been developing (5, 23, ‖) a similar computational theory (see Fig. 1) that attempts to quantitatively account for a host of recent anatomical and physiological data.

The model is a simple and direct extension of the Hubel and Wiesel simple-to-complex cell hierarchy: It takes as an input a gray-value image (256 × 256 pixels, ≈7° × 7° of visual angle) that is first analyzed by a multidimensional array of simple S_1 units which, like cortical simple cells, respond best to oriented bars and edges. S_1 units are modeled as half-rectified filters consisting of aligned and alternating "on" and "off" subregions, which share a common axis of elongation that defines the cell-preferred orientation [see supporting information (SI) *Text* for details]. S_1 units come in four orientations and several different scales (see SI Fig. 9) and densely cover the input image. The next C_1 level corresponds to striate complex cells (1). 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 and sizes is that C_1 units become insensitive to the location and scale of the stimulus within their receptive fields, which is a hallmark of cortical complex cells (1). The parameters of the S_1 and C_1 units (see SI Table 1) were adjusted so as to match as closely as possible the tuning properties of V1 parafoveal simple and complex cells (receptive field size, peak frequency, frequency, and orientation bandwidth; see ref. 24 for details).

Feedforward theories of visual processing, like the model described here, consist of extending these two classes of simple and complex cells to extrastriate areas. By alternating between S layers of simple units and C layers of complex units, the model achieves a difficult tradeoff between selectivity and invariance: Along the hierarchy, at each S stage, simple units become tuned to features of increasing complexity (e.g., from single oriented bars to combinations of oriented bars forming corners and features of intermediate complexities) by combining afferents (C units) with different selectivities (e.g., units tuned to edges at different orientations). For instance, at the S_2 level (respectively S_3), units pool the activities of retinotopically organized afferent C_1 units (respectively C_2 units) with different orientations (different feature tuning), thus increasing the complexity of the representation: From single bars to combinations of oriented bars forming contours or boundary conformations. Conversely, at each C stage, complex units become

Author contributions: T.S., A.O., and T.P. designed research; T.S. and A.O. performed research; T.S. analyzed data; and T.S., A.O., and T.P. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: V1, primary visual cortex; V2, extrastriate visual area II; V4, extrastriate visual area IV; IT, inferotemporal cortex ; PFC, prefrontal cortex; SOA, stimulus onset asynchrony; ISI, interstimulus interval.

[§] To whom correspondence should be addressed. E-mail: serre@mit.edu.

[¶] Thorpe, S., Biologically Motivated Computer Vision, Second International Workshop, Nov. 22–24, 2002, Tübingen, Germany, pp. 1–15.

[‖] Serre, T., Riesenhuber, M., Louie, J., Poggio, T., Biologically Motivated Computer Vision, Second International Workshop, Nov. 22–24, 2002, Tübingen, Germany, pp. 387–397.

This article contains supporting information online at www.pnas.org/cgi/content/full/0700622104/DC1.

www.pnas.org/cgi/doi/10.1073/pnas.0700622104

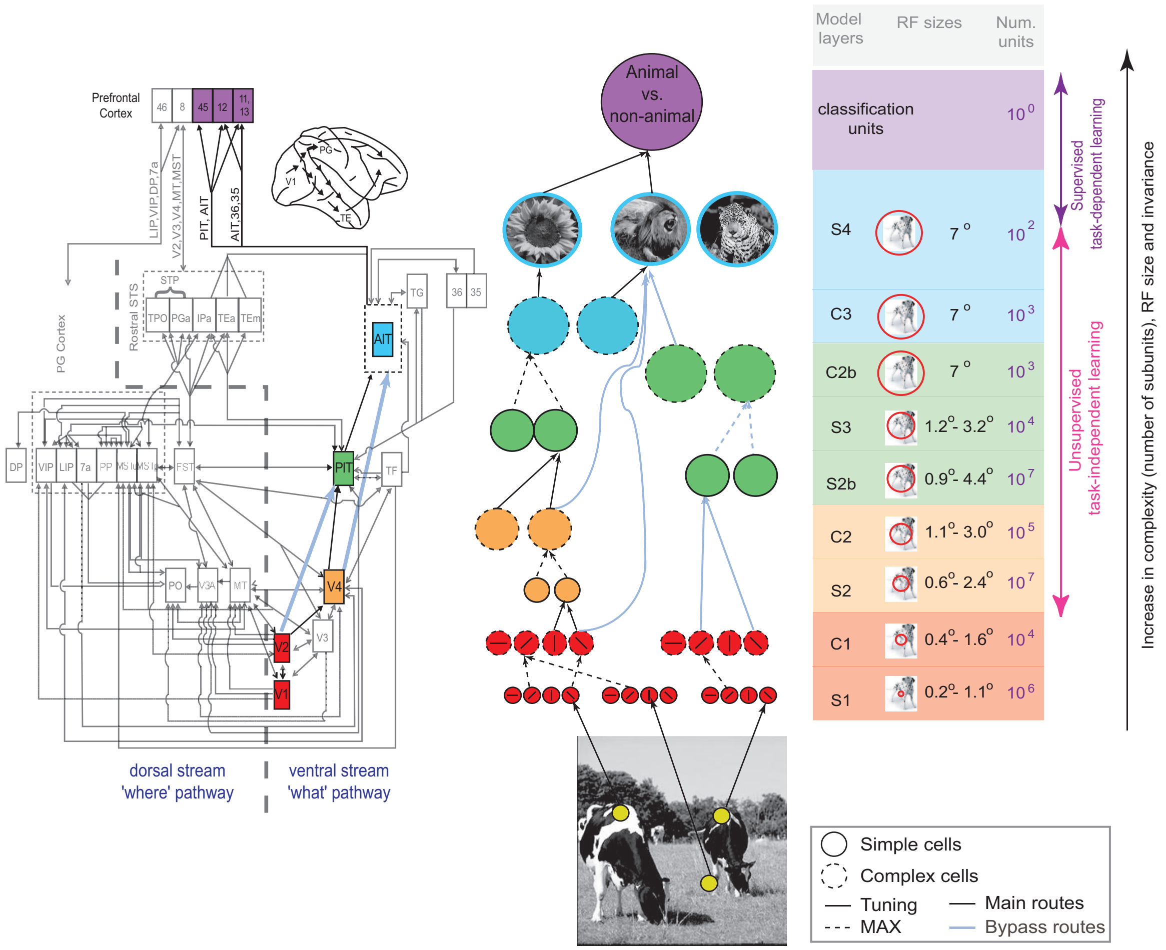


Fig. 1. Sketch of the model. Tentative mapping between the ventral stream in the primate visual system (*Left*) and the functional primitives of the feedforward model (*Right*). The model accounts for a set of basic facts about the cortical mechanisms of recognition that have been established over the last decades: From V1 to IT, there is an increase in invariance to position and scale (1, 2, 4–6), 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, 7). Finally, adult plasticity and learning are probably present at all stages and certainly at the level of IT (6) and PFC. The theory assumes that one of the main functions of the ventral stream, just a part of the visual cortex, is to achieve a tradeoff between selectivity and invariance within a hierarchical architecture. As in ref. 5, 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 operation on their inputs and provide invariance to position and scale (pooling over scales is not shown). The tuning of the S_2, S_{2b}, and S_3 units (corresponding to V2, V4, and the posterior inferotemporal cortex) is determined here by a prior developmental-like unsupervised learning stage (see *SI Text*). Learning of the tuning of the S_4 units and of the synaptic weights from S_4 to the top classification units is the only task-dependent, supervised-learning stage. The main route to IT is denoted with black arrows, and the bypass route (38) is denoted with blue arrows (see *SI Text*). The total number of units in the model simulated in this study is on the order of 10 million. Colors indicate the correspondence between model layers and cortical areas. The table (*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 the visual cortex. Of course, other cortical areas (e.g., in the dorsal stream) as well as noncortical structures (e.g., basal ganglia) are likely to play a role in the process of object recognition. The diagram (*Left*) is modified from ref. 58 (with permission from the author) which represents a juxtaposition of the diagrams of refs. 46 and 59.

increasingly invariant to 2D transformations (position and scale) by combining afferents (*S* units) with the same selectivity (e.g., a vertical bar) but slightly different positions and scales.

The present theory significantly extends an earlier model (5). It follows the same general architecture and computations. The simple *S* units perform a bell-shaped tuning operation over their inputs. That is, the response *y* of a simple unit receiving the pattern of synaptic inputs $(x_1, \ldots, x_{n_{S_k}})$ from the previous layer is given by

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

where σ defines the sharpness of the tuning around the preferred stimulus of the unit corresponding to the weight vector $w = (w_1, \ldots, w_{n_{S_k}})$. That is, the response of the unit is maximal ($y = 1$) when the current pattern of input x matches exactly the synaptic weight vector w and decreases with a bell-shaped tuning profile as the pattern of input becomes more dissimilar. Conversely, 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 $(x_1, \ldots, x_{n_{C_k}})$ from the previous S_k layer:

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

Details about the two key operations can be found in *SI Text* (see also ref. 23).

This class of models seems to be qualitatively and quantitatively consistent with [and in some cases actually predicts (23)] several properties of subpopulations of cells in V1, V4, IT, and PFC (25) as well as fMRI and psychophysical data. For instance, the model predicts (23), at the C_1 and C_2 levels, respectively, the max-like behavior of a subclass of complex cells in V1 (26) and V4 (27). It

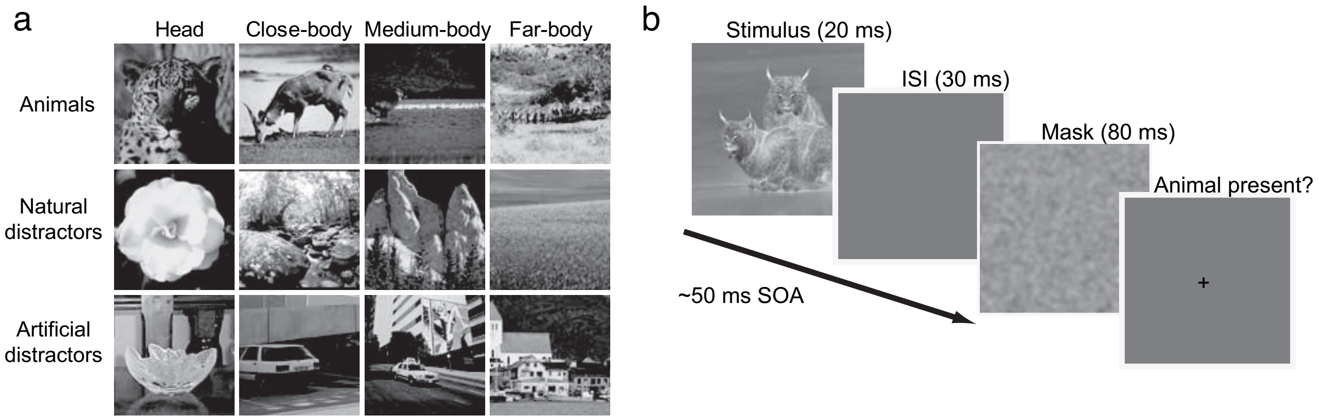


Fig. 2. Animal- vs. non-animal-categorization task. (*a*) The four (balanced) classes of stimuli. Animal images (a subset of the image database used in ref. 30) were manually arranged into four groups (150 images each) based on the distance of the animal from the camera: head (close-up), close-body (animal 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 *Materials and Methods*) was selected so as to prevent human observers and the computational model from relying on low-level cues (see *SI Text*). (*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 an answer of ''yes'' or ''no'' by pressing one of two keys.

also shows good agreement (23) with other data in V4 (28) about the response of neurons to combinations of simple two-bar stimuli (within the receptive field of the S_2 units), and some of the C_2 units in the model show a tuning for boundary conformations which is consistent with recordings from V4 (29) (C. Cadieu, M. Kouh, A. Pasupathy, C. Connor, and T.P., unpublished work). Readout from C_{2b} units in the model described here predicted (23) recent readout experiments in IT (9), showing very similar selectivity and invariance for the same set of stimuli. In addition, plausible biophysical circuits may implement the two key operations (5) assumed by the theory within the time constraints of the experimental data (8).

Because this feedforward model appears to agree with physiological data while performing 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 the visual cortex, and are not implemented in the present feedforward model. Thus, 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 task for which back projections are likely to be inactive (30, 31). A well established experiment is an animal- vs. non-animal-recognition task (30–34).

Results

Animals in natural scenes constitute a challenging class of stimulus because of large variations in shape, pose, size, texture, and position in the scene (see *SI Text* 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 in each set, i.e., 1,200 total stimuli; see *Materials and Methods*), 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. 2*a* and *Materials and Methods*).

When testing human observers, we used a backward-masking protocol (1/f noise image with a duration of 80 ms; see Fig. 2*b*) with a long 50-ms stimulus onset asynchrony [SOA; 50-ms SOA corresponding to a 20-ms stimulus presentation followed by a 30-ms interstimulus interval (ISI)]. It was found (31) that increasing the SOA on a similar animal- vs. non-animal-categorization task above 44 ms only has a minor effect on performance (accuracy scores for longer SOA conditions were not significantly different). At the same time, we expect the mask to block significant top-down effects through the back projections (see *Discussion* and *SI Text*). In the present version of the model, processing by the units (the nodes of the graph in Fig. 1) is approximated as essentially instantaneous (see, however, possible microcircuits involved in the tuning and max operation in ref. 23). All of the processing time would be taken by synaptic latencies and conduction delays (see *SI Text*). The model was compared with human observers in three different experiments.

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. 3*a*. Performance is measured by the d', a monotonic function of the performance of the observers which combines both the hit and false-alarm rates of each observer into one standardized score [see *Materials and Methods*; other accuracy measures such as error rates or hits gave similar results (see *SI Text*)]. The task-specific circuits of the model from IT to PFC were trained for the animal- vs. non-animal-categorization task in a supervised way using a random split procedure (see *Materials and Methods*) on the entire database of stimuli (i.e., in a given run, half of 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 apart from the supervised stage from IT to PFC were fixed before all tests by taking into account the physiology data from V1 to IT). The accuracy of the human observers is well within the range of data previously obtained with go/no-go tasks on similar tasks (30, 31, 33).

Most importantly, 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 values above each panel) who classified it as an animal (irrespective of whether the image contains an animal). For the model, we computed the percentage of times the model (green values) classified

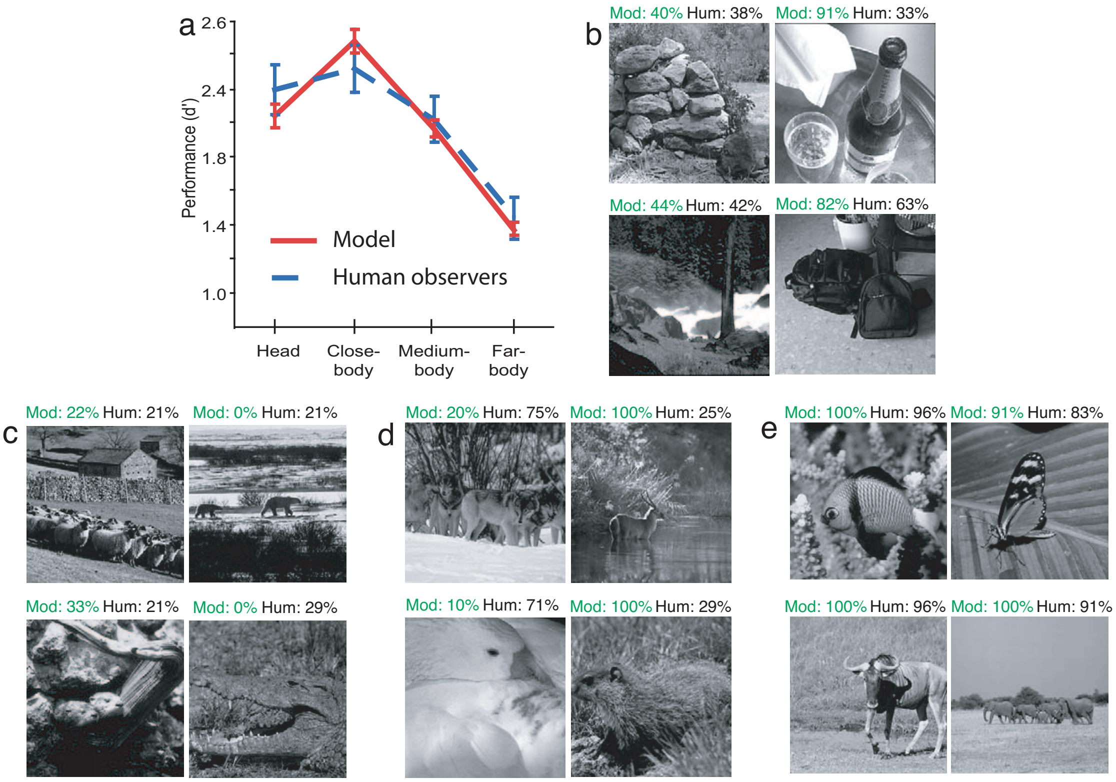


Fig. 3. Comparison between the model and human observers. (*a*) Model- vs. human-level accuracy. Human observers and the model exhibit a very similar pattern of performance (measured with d' measure; see *SI Text*). Error bars indicate the standard errors for the model (computed over $n = 20$ random 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. (*d* and e) Examples of animal images for which the agreement between the model and human observers is poor (*d*) and good (e). The percentages above each thumbnail correspond to the number of times the image was classified as an animal by the model (green values) or by human observers (black values; see *Results* 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).

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. 4), the data suggest that it is the large, overall set of features from V2 to V4 and the posterior inferotemporal cortex that underlies such a human-like performance in this task.

To further test the model, we measured the effect of image rotation (90° and 180°) on performance. Recent behavioral studies (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? SI Fig. 5 shows indeed 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 retrained 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 results (31) to test the influence of the mask on visual processing with four experimental conditions, i.e., when the mask follows the target image (20-ms presentation): (*i*) without any delay ("immediate-mask" condition), (*ii*) with a short ISI of 30 ms (50-ms SOA) as in the previous experiments, (*iii*) with an ISI of 60 ms (80-ms SOA), or (*iv*) 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 SI Fig. 6). 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). The model predicts human-level hit rate very well between the 50- and the 80-ms SOA conditions. For SOAs longer than 80 ms, human observers outperform the model (the performance for the 50-ms SOA condition, however, is only ≈5% lower than the ceiling performance in the no-mask condition). It remains an open question whether the slightly better performance of humans for SOAs longer than 80 ms is due to feedback effects mediated by the back projections (35).

Discussion

The new model implementation used in this study improves the original model (5) in two significant ways. The major extension is a new unsupervised learning stage of the units in intermediate stages of the model (23, ‖). A key assumption in the new model is that the hierarchy of visual areas along the ventral stream of the visual cortex, from V1 to IT, builds a generic dictionary of shape-tuned units which provides a rich representation for task-specific categorization circuits in prefrontal areas. Correspondingly, learn-

**Guyonneau, R., Kirchner, H., Thorpe, S. J., European Conference on Visual Perception, Aug. 22–26, 2005, Coruña, Spain.

ing proceeds in two independent stages: First, during a slow developmental-like unsupervised learning stage, units from V1 to IT become adapted to the statistics of the natural environment (see *SI Text* for details). The resulting dictionary is generic and universal in the sense that it can support several different recognition tasks (23) and in particular, the recognition of many different object categories. After this initial unsupervised learning stage, only the task-specific circuits at the top level in the model, possibly corresponding to categorization units in PFC (25), have to be trained from a small set of labeled examples and in a task-specific manner (see *Materials and Methods*) for the "mature" model to learn a categorization task (e.g., animal vs. non-animal).

Additionally, the new model is closer to the anatomy and the physiology of the visual cortex in terms of quantitative parameter values. For instance, the parameters (see SI Table 1) of the S_1 and C_1 model units were constrained by physiology data (1, 36, 37) so that their tuning properties would agree with those of cortical simple and complex cells (see *SI Text*). In addition to the main routes through the V4 to the IT cortex (4), the model also accounts for the bypass routes (38) from V2 to the posterior inferotemporal cortex and from V4 to the anterior inferotemporal cortex (Fig. 1) [unlike the original model (5)]. A more detailed description of the model can be found in *SI Text*, and a software implementation is accessible from our supplementary online material at http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07/index.htm.

Not only does this class of feedforward models seem to be able to duplicate the tuning properties of at least some cortical cells when probed with artificial stimuli, but it can also handle the recognition of objects in the real world (39, 40) where objects may undergo drastic changes in appearance (e.g., clutter, shape, illumination). Key to the recognition performance of the model is the large number of tuned units across its hierarchical architecture, which is a direct consequence of the learning from natural images and represent a redundant dictionary of fragment-like features (12, 17, 41) 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 10^7 tuned units. 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 (23).

Previous physiological studies have shown that during masked stimulus presentations, the feedforward bottom-up components of cortical cells' response (i.e., the early response from response onset for a period lasting about the stimulus mask) remains essentially unaltered whereas the later response is interrupted (see *SI Text*) (see also refs. 42, 43, and 44 for recent reviews). Several studies (*SI Text*) have shown that this later response includes recurrent processing, that is a modulation through back projections from higher to lower areas. Based on response latencies in the visual cortex (*SI Text*), we estimate that significant top-down modulation should start for stimulus-mask interval ≈40–60 ms (*SI Text*). The model indeed mimics human-level performance for the 50-ms SOA condition. This finding suggests that under these conditions, the present feedforward model may provide a satisfactory description of information processing in the ventral stream of the visual cortex.

Our results indeed agree with several theories of visual processing that suggest that an initial feedforward sweep driven by bottom-up inputs builds a base representation that relies on a basic dictionary of generic features (11–13, 17, 41) before more complex tasks or visual routines can take place through recurrent projections from higher areas (20, 42, 43, 45). Additionally, our results show the limit of what a feedforward architecture can do: In agreement with the human data, the model is able to recognize objects with limited clutter (see ref. 39 for results on a large database of 101 object categories). However, when the amount of clutter present in the images increase, the performance of the model decreases significantly. This suggests a key role for the massive back projections found in the visual cortex (46). Indeed, preliminary results with a simple extension of the present model (47), which requires top-down signals from higher to lower areas to limit visual processing to a "spotlight of attention" centered around the animal target, shows a significant improvement in the classification performance on the "far"-animal condition. In addition, back projections may be important for visual awareness and beyond tasks such as visual categorization for perceptual organization and figure-ground segmentation (48–50) or curve tracing (51).

Nevertheless, our main result is that a simple extension of the feedforward hierarchical architecture, suggested some 40 years ago by Hubel and Wiesel and reflecting the known physiology and anatomy of the visual cortex, correlates well with humans and exhibits comparable accuracy on a difficult (but rapid) recognition task. This finding provides computational neuroscience support to the conjecture that a task-independent, unsupervised, developmental-like learning stage may exist in the ventral stream to generate a large dictionary of shape-tuned units with various degrees of selectivity and invariance from V1 to IT, consistent with recent data (52).

Materials and Methods

Supplementary material is also available at http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07/index.htm and includes, in particular, a basic software implementation for the model, the animal-/non-animal-stimulus database, as well as supplementary data including a summary of different error measures for both the model and human observers (e.g., roc curves).

Stimulus Data Set. All images were gray-value 256 × 256 pixel images. The stimulus database contains a total of 600 animal stimuli (a subset of the Corel database as in ref. 30; 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 being presented in the center of the image) and 600 non-animal stimuli. Animal images were manually grouped into four categories with 150 exemplars in each; that is, 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 (53). We selected images with a mean distance from the camera <1 m for head, between 5 and 20 m for close-body, between 50 and 100 m for medium-body, and >100 m and panoramic views for far-body. The database is publicly available at http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07/index.htm.

Human Psychophysics. All participants (18–35 years old; $n = 24$ in the first experiment with a fixed 50-ms SOA; $n = 14$ in the second experiment with 0°, 90°, and 180° rotated stimuli; $n = 21$ in the last experiment with variable SOAs) gave a written informed consent. 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 (2.4 GHz), 1 GB RAM]. The monitor refresh rate was 100 Hz, allowing stimuli to be displayed with a frame duration of 10 ms and a resolution of 1,024 × 768.

We used MATLAB software (MathWorks, Natick, MA) with the psychophysics toolbox (54, 55) to precisely time the stimulus presentations. In all experiments, the image duration was 20 ms. In all experiments except the last one (see below), the mask appeared after a fixed ISI of 30 ms (corresponding to a SOA of 50 ms). In the last experiment, 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 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, ≈7° × 7° 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 respond as fast and as accurately as possible concerning whether 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 ≈30 min to perform.

Categorization by the Model. To train the PFC classification unit in the model, we used a random splits procedure, which has been shown to provide a good estimate of the expected error of a classifier (56). The procedure was as follows.

1. Split the set of 1,200 (animal and non-animal) images into two halves; denote one half as "training" and the other as "test."
2. Imprint S_4 units with specific examples of animal and non-animal images from the training set of images (25% selected at random). Like units in lower stages become tuned to patches of natural images (see *SI Text*); 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. This finding is consistent with a large body of data that suggests that the selectivity of neurons in IT depends on visual experience (see ref. 23 for a review).
3. Train a PFC classification unit on the labeled "training" set of images. The response y of a classification unit with input weights $c = (c_1, \ldots, c_{K_{S_4}})$, when presented with an input pattern $x = (x_1, \ldots, x_{K_{S_4}})$ from the previous layer (S_4 unit j, denoted x_j, is tuned to the jth training example), is given by

$$y = \sum_j c_j x_j. \qquad [3]$$

The unit response $y \in \mathcal{R}$ is further binarized ($y \lessgtr 0$) to obtain a classification label $\{-1,1\}$. This supervised learning stage involves adjusting the synaptic weights c so as to minimize the overall classification error E on the training set.[††] In this article, 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[‡‡][§§]:

$$E = \sum_{i=1}^{l} \| y^i - \hat{y}^i \|^2 + \lambda \| c \|^2, \qquad [4]$$

where y^i corresponds to the classification unit response for the ith training example, $\hat{y}^i$ is the true label of the ith training example, and λ is a fixed constant. To solve Eq. **1**, we used the nonbiological MATLAB left division operation for matrices, but we obtained similar results with a more biologically plausible stochastic gradient learning approach using weight perturbations modified from ref. 57, i.e., (x^i, y^i) pairs, where x^i denotes the ith image in the training set and y^i is its associated label (animal or non-animal).
4. Evaluate the performance of the classifier on the "test" set. We repeated the overall procedure $n = 20$ times and computed the average model performance. Note that the error bars for the model in Fig. 3 correspond to the standard errors computed over these $n = 20$ random runs.

[††] The full training set is used to adjust the synaptic weights of the classification unit.

[‡‡] Other classifiers could be used (a linear SVM gave very similar results). A recent study (9) demonstrated that a linear classifier can indeed read out with high accuracy and over extremely short time intervals (a single bin as short as 12.5 ms) object identity, object category, and other information (such as the position and size of the object) from the activity of ≈100 neurons in IT.

[§§] A single classifier was trained on all four animal and non-animal categories together.

We thank C. Cadieu, B. Desimone, M. Riesenhuber, U. Knoblich, M. Kouh, G. Kreiman, S. Thorpe, and A. Torralba for comments and fruitful discussions related to this work and S. Das, J. Dicarlo, M. Greene, E. Meyers, E. Miller, P. Sinha, C. Tomasi, and R. VanRullen for comments on the manuscript. This work was sponsored by Defense Advanced Research Projects Agency Contracts HR0011-04-1-0037 and FA8650-06-7632, National Institutes of Health–National Science Foundation Contracts EIA-0218506 and 0546262 (to A.O.), and National Institutes of Health Contract 1 P20 MH66239-01A1. Additional support was provided by the Central Research Institute of Electric Power Industry, Daimler–Chrysler AG, Eastman Kodak Company, Honda Reasearch Institute USA, Komatsu, Merril–Lynch, the NEC Fund, Oxygen, Siemens Corporate Research, Sony, Sumitomo Metal Industries, Toyota Motor Corporation, and the Eugene McDermott Foundation (to T.P.).

1. Hubel DH, Wiesel TN (1968) *J Phys* 195:215–243.
2. Perrett D, Oram M (1993) *Image Vision Comput* 11:317–333.
3. Kobatake E, Tanaka K (1994) *J Neurophysiol* 71:856–867.
4. Tanaka K (1996) *Annu Rev Neurosci* 19:109–139.
5. Riesenhuber M, Poggio T (1999) *Nat Neurosci* 2:1019–1025.
6. Logothetis NK, Pauls J, Poggio T (1995) *Curr Biol* 5:552–563.
7. Desimone R (1991) *J Cognit Neurosci* 3:1–8.
8. Perrett D, Hietanen J, Oram M, Benson P (1992) *Philos Trans R Soc London B* 335:23–30.
9. Hung C, Kreiman G, Poggio T, DiCarlo J (2005) *Science* 310:863–866.
10. Potter M (1975) *Science* 187:565–566.
11. Treisman AM, Gelade G (1980) *Cognit Psychol* 12:97–136.
12. Wolfe J, Bennett S (1997) *Vision Res* 37:25–44.
13. Schyns P, Oliva A (1994) *Psychol Sci* 5:195–200.
14. Fukushima K (1980) *Biol Cybern* 36:193–202.
15. Wallis G, Rolls ET (1997) *Prog Neurobiol* 51:167–194.
16. Mel BW (1997) *Neural Comput* 9:777–804.
17. Ullman S, Vidal-Naquet M, Sali E (2002) *Nat Neurosci* 5:682–687.
18. Amit Y, Mascaro M (2003) *Vision Res* 43:2073–2088.
19. Wersing H, Koerner E (2003) *Neural Comput* 15:1559–1588.
20. Hochstein S, Ahissar M (2002) *Neuron* 36:791–804.
21. Biederman I (1987) *Psychol Rev* 94:115–147.
22. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) *Proc IEEE* 86:2278–2324.
23. Serre T, Kouh, M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005) *MIT AI Memo 2005–036/CBCL Memo 259*, available at ftp://publications.ai.mit.edu/ai-publications/2005/AIM-2005–036.pdf, published online December 19, 2005.
24. Serre T, Riesenhuber M (2004) *MIT AI Memo 2004–017/CBCL Memo 239*, available at ftp://publications.ai.mit.edu/ai-publications/2004/AIM-2004–017.pdf, published online July 27, 2004.
25. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) *Science* 291:312–316.
26. Lampl I, Ferster D, Poggio T, Riesenhuber M (2004) *J Neurophysiol* 92:2704–2713.
27. Gawne TJ, Martin JM (2002) *J Neurophysiol* 88:1128–1135.
28. Reynolds JH, Chelazzi L, Desimone R (1999) *J Neurosci* 19:1736–1753.
29. Pasupathy A, Connor CE (2001) *J Neurophysiol* 86:2505–2519.
30. Thorpe S, Fize D, Marlot C (1996) *Nature* 381:520–522.
31. Bacon-Mace N, Mace M, Fabre-Thorpe M, Thorpe S (2005) *Vision Res* 45:1459–1469.
32. Thorpe S, Fabre-Thorpe M (2001) *Science* 291:260–263.
33. VanRullen R, Koch C (2003) *J Cognit Neurosci* 15:209–217.
34. Rousselet G, Mace M, Fabre-Thorpe M (2003) *J Vision* 3:440–455.
35. Bienenstock E, Geman S, Potter D (1997) *Advances in Neural Information Processing Systems*, eds Mozer MC, Jordan MI, Petsche T (MIT Press, Cambridge, MA), pp 838–834.
36. Schiller PH, Finlay BL, Volman SF (1976) *J Neurophysiol* 39:1288–1319.
37. DeValois R, Albrecht D, Thorell L (1982) *Vision Res* 22:545–559.
38. Nakamura H, Gattass R, Desimone R, Ungerleider LG (1993) *J Neurosci* 13:3681–3691.
39. Serre T, Wolf L, Bileschi S, Poggio T (2007) *IEEE Trans Pattern Anal Mach Intell* 29:411–426.
40. Serre T, Wolf L, Poggio T (2005) *Proc IEEE Conf Comput Vision Pattern Recognit* 2:994–1000.
41. Evans K, Treisman A (2005) J Exp Psychol Hum Percept Perform 31:1476–1492.
42. Enns J, Lollo VD (2000) *Trends Cognit Neurosci* 4:345–351.
43. Lamme V, Roelfsema P (2000) *Trends Neurosci* 23:571–579.
44. Breitmeyer B, Ogmen H (2006) *Visual Masking: Time Slices Through Conscious and Unconscious Vision* (Oxford Univ Press, Oxford).
45. Roelfsema P, Lamme V, Spekreijse H (2000) *Vision Res* 40:1385–1411.
46. Felleman DJ, van Essen DC (1991) *Cereb Cortex* 1:1–47.
47. Serre T (2006) PhD thesis (MIT, Cambridge, MA).
48. Roelfsema P, Lamme V, Spekreijse H, Bosch H (2002) *J Cognit Neurosci* 12:525–537.
49. Lamme V, Zipser K, Spekreijse H (2002) *J Cognit Neurosci* 14:1044–1053.
50. Lee T, Mumford D (2003) *J Opt Soc Am* 20:1434–1448.
51. Roelfsema PR, Lamme VA, Spekreijse H (1998) *Nature* 395:376–381.
52. Freedman D, Riesenhuber M, Poggio T, Miller E (2006) *Cereb Cortex*, in press.
53. Torralba A, Oliva A (2002) *IEEE Trans Pattern Anal Mach Intell* 24:1226–1238.
54. Brainard D (1997) *Spat Vis* 10:433–436.
55. Pelli D (1997) *Spat Vis* 10:437–442.
56. Devroye L, Laszlo G, Lugosi G (1996) *A Probabilistic Theory of Pattern Recognition* (Springer, New York).
57. Sutton R, Barto A (1981) *Psychol Rev* 88:135–170.
58. Gross CG (1998) *Brain Vision and Memory: Tales in the History of Neuroscience* (MIT Press, Cambridge, MA).
59. Ungerleider L, Mishkin M (1982) in *Analysis of Visual Behavior*, eds Ingle DJ, Goodale MA, Mansfield RJW (MIT Press, Cambridge, MA), pp 549–586.

SI Text

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

Model Overview

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

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

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

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

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

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

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

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

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

S_1 units

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

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

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

The five parameters, that is, the orientation θ, the aspect ratio γ, the effective width σ, the phase ϕ and the wavelength λ, determine the properties of the spatial receptive field of the S_1 units. In setting these parameters we tried to generate a population of units that match the bulk of parafoveal cells as closely as possible (see ref. [12]).

C_1 units

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

S_2 units

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

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

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

C_2 units

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

S_3 and C_3 units

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

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

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

Unsupervised Learning from V2 to IT

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

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

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

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

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

The d' sensitivity measure

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

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

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

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

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

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

On interrupting recurrent processing with the mask

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

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

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

Benchmarking the database

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

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

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

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

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

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

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

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

25. Heisele, B, Serre, T, Pontil, M, Vetter, T, & Poggio, T. (2002) *Advances in Neural Information Processing Systems* **14**, 1239–1245.

26. Fei-Fei, L, Fergus, R, & Perona, P. (2004) *Proc. IEEE CVPR, Workshop on Generative-Model Based Vision*.

27. Ullman, S, Vidal-Naquet, M, & Sali, E. (2002) *Nat. Neurosci.* **5**, 682–687.

28. Treisman, A. M & Gelade, G. (1980) *Cog. Psych.* **12**, 97–136.

29. Evans, K & Treisman, A. (2005) *J. Exp. Psych.: Hum. Percept. Perf.* **31**, 1476–1492.

30. Wolfe, J & Bennett, S. (1997) *Vis. Res.* **37**, 25–44.

31. Markram, H, Lübke, J, Frotscher, M, & Sakmann, B. (1997) *Science* **275**, 213–215.

32. Bi, G & Poo, M. (1998) *J. Neurosci.* **18**, 10464–10472.

33. Abarbanel, H, Huerta, R, & Rabinovich, M. (2002) *Proc. Nat. Acad. Sci. USA* **99**, 10132–10137.

34. van Rossum, M, Bi, G, & Turrigiano, G. (2000) *J. Neurosci.* **20**, 8812–8821.

35. Földiák, P. (1991) *Neural Comp.* **3**, 194–200.

36. Macmillan, N. A & Creelman, C. D. (1991) *Detection Theory: A User's Guide*. (Cambridge University Press).

37. Breitmeyer, B & Ogmen, H. (2006) *Visual Masking: Time Slices through Conscious and Unconscious Vision*. (Oxford University Press).

38. Bridgeman, B. (1980) *Brain Res.* **196**, 347–364.

39. Macknik, S & Livingstone, M. (1998) *Nat. Neurosci.* **1**, 144–149.

40. Lamme, V, Zipser, K, & Spekreijse, H. (2002) *J. Cogn. Neurosci.* **14**, 1044–1053.

41. Kovács, G, Vogels, R, & Orban, G. (1995) *Proc. Nat. Acad. Sci. USA* **92**, 5587–5591.

42. Rolls, E, Tovee, M, & Panzeri, S. (1999) *J. Cogn. Neurosci.* **11**, 300–311.

43. Keysers, C, Xiao, D. K, Földiák, P, & Perrett, D. I. (2001) *J. Cogn. Neurosci.* **13**, 90–101.

44. Thompson, K & Schall, J. (1999) *Nat. Neurosci.* **2**, 283–288.

45. Lamme, V & Roelfsema, P. (2000) *Trends in Neurosci.* **23**, 571–579.

46. Enns, J & Lollo, V. D. (2000) *Trends in Cogn. Sci.* **4**, 345–351.

47. Nowak, L & Bullier, J. (1997) *Extrastriate visual cortex in primates*. (New York: Plenum Press) Vol. 12, pp. 205–241.

48. Thorpe, S & Fabre-Thorpe, M. (2001) *Science* **291**, 260–263.

49. Knierim, J & van Essen, D. (1992) *J. Neurophys.* **67**, 961–p80.

50. Zhou, H, Friedman, H. S, & von der Heydt, R. (2000) *J. Neurosci.* **20**, 6594–6611.

51. Corthout, E, Uttl, B, Walsh, V, Hallett, M, & Cowey, A. (1999) *Neuroreport*.

52. Hochstein, S & Ahissar, M. (2002) *Neuron* **36**, 791–804.

53. Renninger, L & Malik, J. (2004) *Vis. Res.* **44**, 2301–2311.

54. Torralba, A & Oliva, A. (2003) *Network: computation in neural systems* **14**, 391–412.

55. Kobatake, E, Wang, G, & Tanaka, K. (1998) *J. Neurophys.* **80**, 324–330.

Table 1. Summary of all model parameters

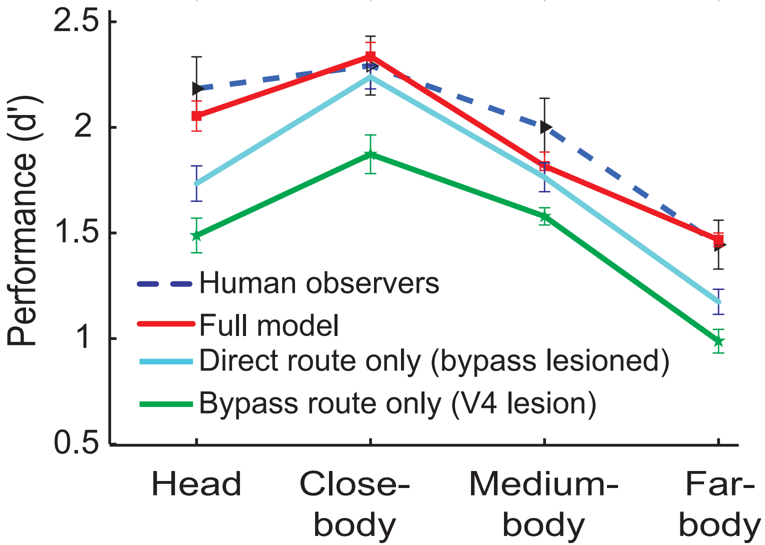
	S_1 parameters							
RF size (pix.)	7 ; 9	11 ; 13	15 ; 17	19 ; 21	23 ; 25	27 ; 29	31 ; 33	35 ; 37 ; 39
σ	2.8 ; 3.6	4.5 ; 5.4	6.3 ; 7.3	8.2 ; 9.2	10.2 ; 11.3	12.3 ; 13.4	14.6 ; 15.8	17.0 ; 18.2 ; 19.5
λ	3.5 ; 4.6	5.6 ; 6.8	7.9 ; 9.1	10.3 ; 11.5	12.7 ; 14.1	15.4 ; 16.8	18.2 ; 19.7	21.2 ; 22.8 ; 24.4
θ	$0^0; 45^0; 90^0; 180^0$							
No. S_1 types	$K_{S_1} = 4$							
	C_1 parameters							
Bands ΔS_{C_1}	1	2	3	4	5	6	7	8
Grid size $\Delta N^S_{C_1}$	8	10	12	14	16	18	20	22
Sampling ϵ_{C_1}	3	5	7	8	10	12	13	15
No. C_1 types	$K_{C_1} = K_{S_1} = 4$							
	S_2 parameters							
Grid size ΔN_{S_2}	3×3 ($\times 4$ orientations)							
No inp. n_{S_2}	10							
No. S_2 types	$K_{S_2} \approx 2000$							
	C_2 parameters							
Bands ΔS_{C_2}	1 ; 2		3 ; 4		5 ; 6		7 ; 8	
Grid size $\Delta N^S_{C_2}$	8		12		16		20	
Sampling ϵ_{C_2}	3		7		10		13	
No. C_2 types	$K_{C_2} = K_{S_2} \approx 2000$							
	S_3 parameters							
Grid size ΔN_{S_3}	3×3 ($\times K_{S_2}$)							
No. inp. n_{S_3}	100							
No. S_3 types	$K_{S_3} \approx 2000$							
	C_3 parameters							
Bands ΔS_{C_3}	1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8							
Grid size $\Delta N^S_{C_3}$	40							
No. C_3 types	$K_{C_3} = K_{S_3} \approx 2000$							
	S_{2b} parameters							
Grid size $\Delta N_{S_{2b}}$	$6 \times 6; 9 \times 9; 12 \times 12; 15 \times 15$ ($\times 4$ orientations)							
No. inp. $n_{S_{2b}}$	100							
No. S_{2b} types	$K_{S_{2b}} \approx 500$ for each size ≈ 2000 total							
	C_{2b} parameters							
Bands $\Delta S_{C_{2b}}$	1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8							
Grid size $\Delta N^S_{C_{2b}}$	40							
No. C_{2b} types	$K_{C_{2b}} = K_{S_{2b}} \approx 500$ for each size≈ 2000 total							

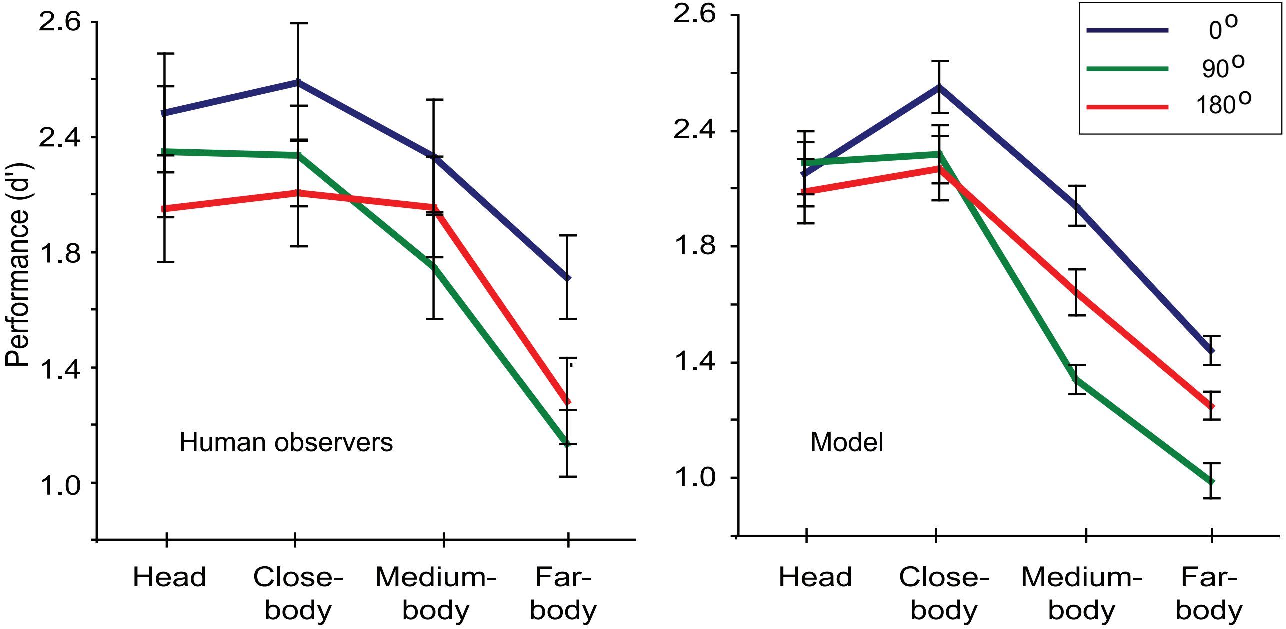
See supporting information text for details.

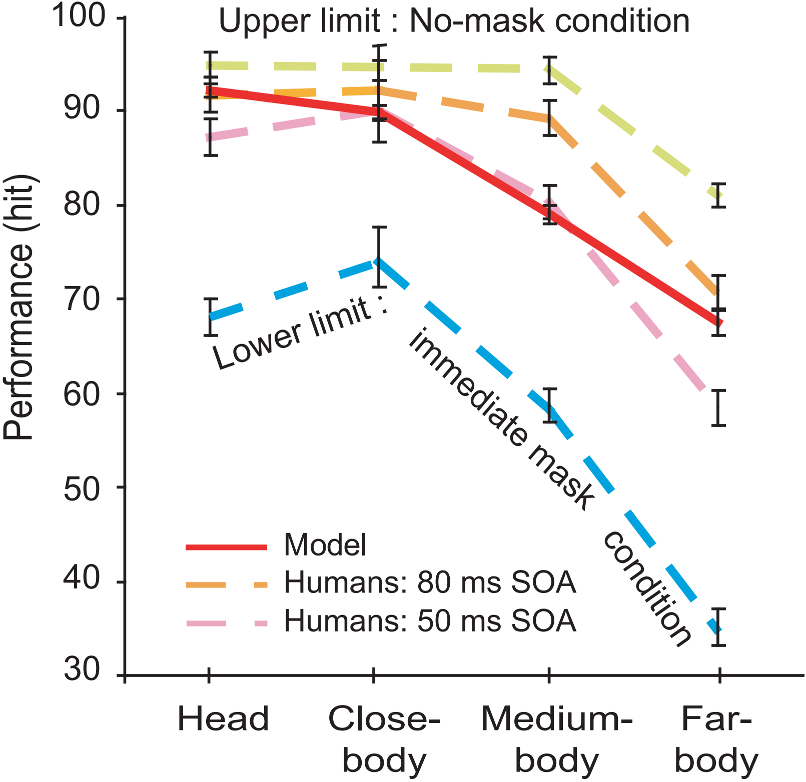
Table 2. Summary of the performance of several benchmark computer vision systems on the animal database

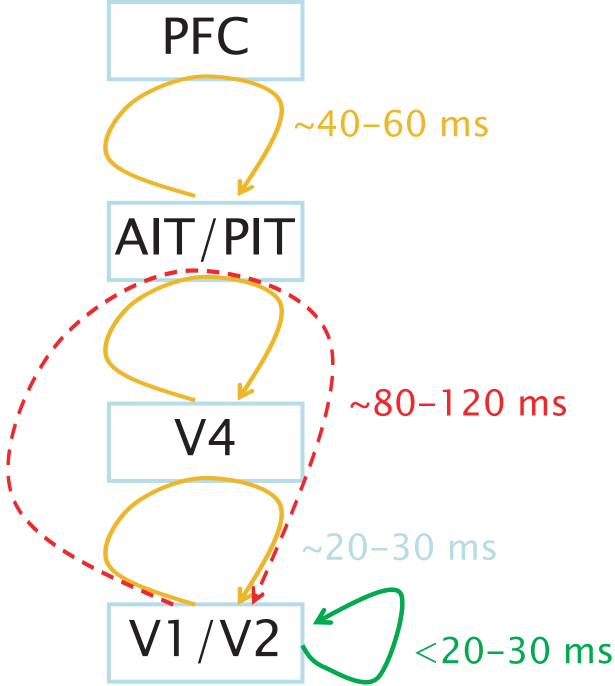
	Head	Close body	Medium body	Far body
Mean luminance	0.28	0.36	0.46	0.34
Gray value SVM	0.23	0.22	0.17	0.13
Textons	0.84	0.58	0.69	0.35
Global features	1.43	1.73	1.47	0.74
Model C_1 layer	1.37	1.78	1.53	0.65

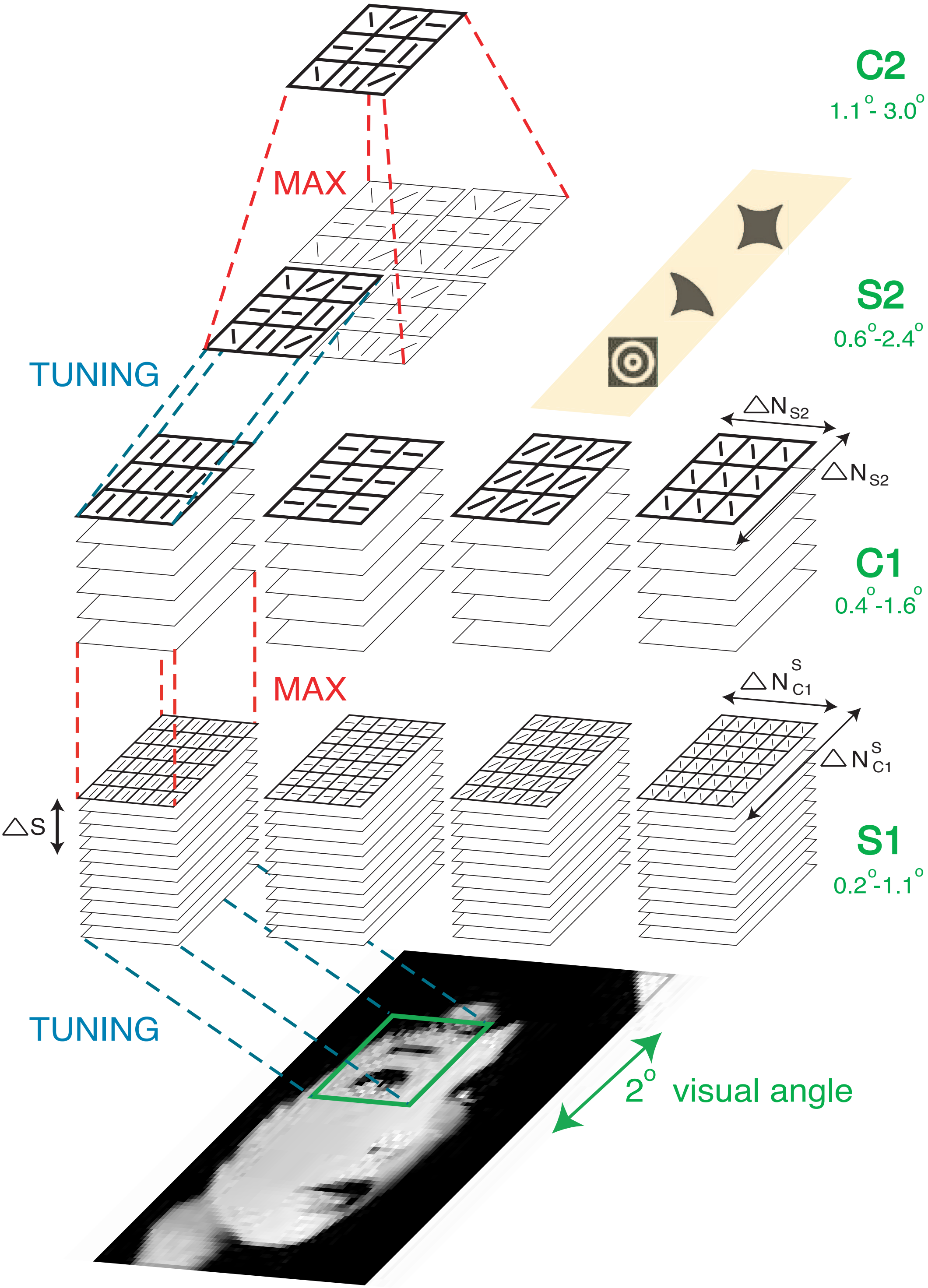
The poor performance of simple classification strategies indicate that it is very unlikely that human observers could rely on low-level cues. The performance of the model is higher than all of the benchmarks with a d' of 2.04, 2.48, 1.97, 1.37 on the subcategories.


Performance (d')
2.5
2
1.5
1
0.5
Human observers
Full model
Direct route only (bypass lesioned)
Bypass route only (V4 lesion)
Head
Close-body
Medium-body
Far-body


2.6
2.4
1.8
1.4
1.0
Performance (d')
Human observers
Head
Close-body
Medium-body
Far-body
2.6
2.4
1.8
1.4
1.0
Model
Head
Close-body
Medium-body
Far-body
0°
90°
180°


Upper limit : No-mask condition
Lower limit : immediate mask condition
Performance (hit)
100
90
80
70
60
50
40
30
Head
Close-body
Medium-body
Far-body
Model
Humans: 80 ms SOA
Humans: 50 ms SOA


PFC
~40–60 ms
AIT / PIT
~80–120 ms
V4
~20–30 ms
V1 / V2
<20–30 ms



C2
1.1° - 3.0°
MAX
S2
0.6°-2.4°
TUNING
ΔN S2
ΔN S2
C1
0.4°-1.6°
MAX
ΔN S C1
ΔN S C1
ΔS
S1
0.2°-1.1°
TUNING
2° visual angle