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,[§]Department of Brain and Cognitive Sciences, and [¶]Massachusetts Institute of Technology

Submitted to Proceedings of the National Academy of Sciences of the United States of America

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 processing 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 & 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, ventral stream, visual cortex, natural scenes, rapid categorization, feedforward architecture, features of intermediate complexity, computational neuroscience

Object recognition in cortex is mediated by the ventral visual pathway running from V1 [1] through extrastriate visual areas V2 and V4 to IT [2, 3, 4], and then to PFC which is involved in linking perception to memory and action. Over the last decade, a number of physiological studies in non-human 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 [5, 1, 2, 6, 4] 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, 7, 3]. 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 visual cortex, remains a matter of debate. 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 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 about 100 ms after stimulus onset, contains surprisingly accurate and robust information supporting a variety of recognition tasks. While this 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, 11, 12, 13].

One of the first feedforward models, Fukushima's Neocognitron [14], followed the basic Hubel & 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 upon several existing neurobiological models [15, 16, 5, 17, 18, 19, 20], conceptual proposals [1, 2, 21, 22] and computer vision systems [14, 23], we have been developing [5, 24, 25] 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 & Wiesel simple-to-complex cell hierarchy: it takes as an input a gray-value image (256×256 pixels $\sim 7^o \times 7^o$ of visual angle) that is first analyzed by a multi-dimensional 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 *SI* for details). S_1 units come in four orientations and several different scales (see Fig. *SI* 6) 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 Table *SI* 1) were adjusted so as to match as closely as possible the tuning properties of V1 parafoveal simple and complex cells (RF size, peak frequency, frequency, and orientation bandwidth, see [26] for details).

Feedforward theories of visual processing, like the model described here, consist in 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 trade-off 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 to form 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

Conflict of interest footnote placeholder

Abbreviations: V1, primary visual cortex; V2, visual area II; V4, visual area IV; PIT, posterior inferotemporal cortex; AIT, anterior inferotemporal cortex; PFC, prefontal cortex

* To whom correspondence should be addressed. E-mail: serre@mit.edu

stage, complex units become 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 signficantly extends an earlier model [5]. It follows the same general architecture and computations. The *simple* S units perform a bell-shape TUNING operation over their inputs. 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:

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

where σ defines the sharpness of the TUNING around the preferred stimulus of the unit corresponding to the weight vector $\mathbf{w} = (w_1, \ldots w_{n_{S_k}})$. 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. 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 $\left(x_1, \ldots, x_{n_{C_k}}\right)$ from the previous S_k layer:

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

Details about these the two key operations can be found in the *SI* (see also [25]).

This class of models seems to be qualitatively and quantitatively consistent with (and in some cases actually predicts, see [25]) several properties of subpopulations of cells in V1, V4, IT, and PFC [27] as well as fMRI and psychophysical data. For instance, the model predicts [25], at the C_1 and C_2 levels respectively, the max-like behavior of a subclass of complex cells in V1 [28] and V4 [29]. It also shows good agreement [25] with other data in V4 [30] 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 [31] (see also (Cadieu, Kouh, Pasupathy, Connor et al, in prep). Read-out from C_{2b} units in the model described here predicted [25] recent read-out 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 visual cortex and 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 [32, 33]. A well-established experiment is an animal vs. non-animal recognition task [32, 34, 35, 36, 33].

Results

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* 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 *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. 2a, 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. 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). It was found [33] 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 later and *SI*). 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 [25]). All the processing time would be taken by synaptic latencies and conduction delays (see *SI*). The model was compared to 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. 3a. 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*). 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 *Materials and Methods*) 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 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 [32, 35, 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 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 value $p < 0.01$). Together with the results of a "lesion study" performed on the model (see Fig. *SI* 1) the

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

To further test the model we measured the effect of image rotation (90^o and 180^o) on performance. Recent behavioral studies [36] (see also abstract by *Guyonneau, Kirchner and Thorpe, ECVP 2005*) 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. *SI* 2 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 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 results [33] 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) as in the previous experiments; c) with an ISI of 60 ms (80 ms SOA) 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. *SI* 3). 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 ms SOA 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 about 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 [37].

Discussion

The new model implementation used in this paper improves the orginal model [5] in two significant ways. The major extension is a new unsupervised learning stage of the units in intermediate stages of the model [24, 25]. 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, learning 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* for details). The resulting dictionary is generic and universal, in the sense that it can support several different recognition tasks [25] and in particular the recognition of many different object categories. After this initial unsupervised learning stage, for the "mature" model to learn a categorization task (*e.g.*, animal vs. non-animal) only the task-specific circuits at the top level in the model, possibly corresponding to categorization units in PFC [27], have to be trained from a small set of labeled examples and in a task specific manner (see *Materials and Methods*).

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 Table *SI* 1) of the S_1 and C_1 model units were constrained by physiology data [1, 38, 39] so that their tuning properties would agree with those of cortical simple and complex cells (see *SI*). In addition to the main routes through V4 to IT cortex [4] the model also accounts for the bypass routes [40] from V2 to PIT and V4 to AIT (see Fig. 1; unlike the original model [5]). A more detailed description of the model can be found in *SI* 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 [41, 42] 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 [17, 43, 12] 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 (see [25]).

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 of time lasting about the stimulus-mask) remains essentially unaltered while the later response is interrupted (see *SI* and [44, 45, 46] for recent reviews). Several studies (see *SI*) 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 (see *SI*), we estimate that significant top-down modulation should start for stimulus-mask interval around 40-60 ms (see *SI*). The model indeed mimics human-level performance for the 50 ms SOA condition. This suggests that, under these conditions, the present feedforward model may provide a satisfactory description of information processing in the ventral stream of 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, 12, 13, 17, 43] before more complex tasks or visual routines can take place through recurrent projections from higher areas [47, 44, 45, 21]. 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 also [41] 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 [48]. Indeed, preliminary results with a simple extension of the present model (see [49]) 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 [50, 51, 52] or curve tracing [53].

Nevertheless, our main result is that a simple extension of the feedforward hierarchical architecture, suggested some forty years ago by Hubel & Wiesel and reflecting the known physiology and anatomy of visual cortex, correlates well with humans and exhibit comparable accuracy on a difficult (but rapid) recognition task. This 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, consistently with recent data [54].

Materials and methods

Supplementary web material is also available (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).

The stimulus dataset. 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 [32]; 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) 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 [55]. 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://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07/index.htm.

Human psychophysics. All participants (18 to 35 years of age; $n = 24$ in the first experiment with a fixed 50 ms SOA; $n = 14$ in the second experiment with 0^o, 90^o and 180^o rotated stimuli; $n = 21$ in the last experiment with a 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, 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 [56, 57] 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 inter-stimulus interval (ISI) of 30 ms (corresponding to a Stimulus Onset Asynchrony 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, about $7^o \times 7^o$ 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 approximately 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 [58]. The procedure is as follow:

1) *Split* the set of 1,200 (animal and non-animal) images into two halves; denote one half *Training* and the other *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*), 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 is, consistent with a large body of data that suggests that the selectivity of neurons in IT depends on visual experience (see [25] 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 $\mathbf{c} = \left(c_1, \ldots, c_{K_{S_4}}\right)$, when presented with an input pattern $\mathbf{x} = (x_1, \ldots x_{K_{S_4}})$ from the previous layer (S_4 unit j, denoted x_j, is tuned to the j^{th} 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 $\mathbf{c}$ so as to minimize the overall classification error E on the training set.[1] 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: [2][3]

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

where y^i corresponds to the classification unit response for the 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. 4 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 [59]. *i.e.*, (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).

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 and 4 in the main manuscript correspond to the standard errors computed over these $n = 20$ random runs.

We are grateful to 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. We would also like to thank S. Das, J. Dicarlo, 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.).

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

[2] 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 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.

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

1. Hubel, D. H & Wiesel, T. N. (1968) *J. Phys.* **195**, 215–243.
2. Perrett, D & Oram, M. (1993) *Img. Vis. Comput.* **11**, 317–333.
3. Kobatake, E & Tanaka, K. (1994) *J. Neurophys.* **71**, 856–867.
4. Tanaka, K. (1996) *Ann. Rev. Neurosci.* **19**, 109–139.
5. Riesenhuber, M & Poggio, T. (1999) *Nat. Neurosci.* **2**, 1019–1025.
6. Logothetis, N. K, Pauls, J, & Poggio, T. (1995) *Curr. Biol.* **5**, 552–563.
7. Desimone, R. (1991) *J. Cogn. Neurosci.* **3**, 1–8.
8. Perrett, D, Hietanen, J, Oram, M, & Benson, P. (1992) *Philos. Trans. Roy. Soc. 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, A. M & Gelade, G. (1980) *Cognit. Psychol.* **12**, 97–136.
12. Wolfe, J & Bennett, S. (1997) *Vis. Res.* **37**, 25–44.
13. Schyns, P & Oliva, A. (1997) *Psychol Sci* **5**, 195–200.
14. Fukushima, K. (1980) *Biol. Cyb.* **36**, 193–202.
15. Wallis, G & Rolls, E. T. (1997) *Prog. Neurobiol.* **51**, 167–194.
16. Mel, B. W. (1997) *Neural Comp.* **9**, 777–804.
17. Ullman, S, Vidal-Naquet, M, & Sali, E. (2002) *Nat. Neurosci.* **5**, 682–687.
18. Thorpe, S. (2002) *Biologically Motivated Computer Vision, Second International Workshop (BMCV 2002)* pp. 1–15.
19. Amit, Y & Mascaro, M. (2003) *Vis. Res.* **43**, 2073–2088.
20. Wersing, H & Koerner, E. (2003) *Neural Comp.* **15**, 1559–1588.
21. Hochstein, S & Ahissar, M. (2002) *Neuron* **36**, 791–804.
22. Biederman, I. (1987) *Psych. Rev.* **94**, 115–147.
23. LeCun, Y, Bottou, L, Bengio, Y, & Haffner, P. (1998) *Proc. of the IEEE* **86**, 2278–2324.
24. Serre, T, Riesenhuber, M, Louie, J, & Poggio, T. (2002) *Biologically Motivated Computer Vision, Second International Workshop (BMCV 2002)* pp. 387–397.
25. Serre, T, Kouh., M, Cadieu, C, Knoblich, U, Kreiman, G, & Poggio, T. (2005) *MIT AI Memo 2005-036 / CBCL Memo 259*.
26. Serre, T & Riesenhuber, M. (2004) *MIT AI Memo 2004-017 / CBCL Memo 239*.
27. Freedman, D. J, Riesenhuber, M, Poggio, T, & Miller, E. K. (2001) *Science* **291**, 312–316.
28. Lampl, I, Ferster, D, Poggio, T, & Riesenhuber, M. (2004) *J. Neurophys.* **92**, 2704–2713.
29. Gawne, T. J & Martin, J. M. (2002) *J. Neurophys.* **88**, 1128–1135.
30. Reynolds, J. H, Chelazzi, L, & Desimone, R. (1999) *J. Neurosci.* **19**, 1736–1753.
31. Pasupathy, A & Connor, C. E. (2001) *J. Neurophys.* **86**, 2505–2519.
32. Thorpe, S, Fize, D, & Marlot, C. (1996) *Nature* **381**, 520–522.
33. Bacon-Mace, N, Mace, M, Fabre-Thorpe, M, & Thorpe, S. (2005) *Vis. Res.* **45**, 1459–1469.
34. Thorpe, S & Fabre-Thorpe, M. (2001) *Science* **291**, 260–263.
35. VanRullen, R & Koch, C. (2003) *J. Cogn. Neurosci.* **15**, 209–217.
36. Rousselet, G, Mace, M, & Fabre-Thorpe, M. (2003) *Journal of Vision* **3**, 440–455.
37. Bienenstock, E, Geman, S, & Potter, D. (1997) *Advances in Neural Information Processing Systems* pp. 838–834.
38. Schiller, P. H, Finlay, B. L, & Volman, S. F. (1976) *J. Neurophysiol.* **39**, 1288–1319.
39. DeValois, R, Albrecht, D, & Thorell, L. (1982) *Vis. Res.* **22**, 545–559.
40. Nakamura, H, Gattass, R, Desimone, R, & Ungerleider, L. G. (1993) *J. Neurosci.* **13**, 3681–3691.
41. Serre, T, Wolf, L, Bileschi, S, & Poggio, T. (2007) *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**, 411–426.
42. Serre, T, Wolf, L, & Poggio, T. (2005) *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* **2**, 994–1000.
43. Evans, K & Treisman, A. (2005) *J. Exp. Psych.: Hum. Percept. Perf.* **31**, 1476–1492.
44. Enns, J & Lollo, V. D. (2000) *Theoretical Computer Science* **4**, 345–351.
45. Lamme, V & Roelfsema, P. (2000) *Trends in Neurosci.* **23**, 571–579.
46. Breitmeyer, B & Ogmen, H. (2006) *Visual Masking: Time Slices through Conscious and Unconscious Vision*. (Oxford University Press).
47. Roelfsema, P, Lamme, V, & Spekreijse, H. (2000) *Vis. Res.* **40**, 1385–1411.
48. Felleman, D. J & van Essen, D. C. (1991) *Cereb. Cortex* **1**, 1–47.
49. Serre, T. (2006) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge, MA).
50. Roelfsema, P, Lamme, V, Spekreijse, H, & Bosch, H. (2002) *J. Cogn. Neurosci.* **12**, 525–537.
51. Lamme, V, Zipser, K, & Spekreijse, H. (2002) *J. Cogn. Neurosci.* **14**, 1044–1053.
52. Lee, T & Mumford, D. (2003) *Journal of the Optical Society of America* **20**, 1434–1448.
53. Roelfsema, P. R, Lamme, V. A, & Spekreijse, H. (1998) *Nature* **395**, 376–381.
54. Freedman, D, Riesenhuber, M, Poggio, T, & Miller, E. (2006) *Cereb. Cortex*. in press.
55. Torralba, A & Oliva, A. (2002) *IEEE Pattern Analysis and Machine Intelligence* **24**, 1226–1238.
56. Brainard, D. (1997) *Spat. Vis.* **10**, 433–436.
57. Pelli, D. (1997) *Spat. Vis.* **10**, 437–442.
58. Devroye, L, Laszlo, G, & Lugosi, G. (1996) *A probabilistic theory of pattern recognition*. (Springer-Verlag, New York).
59. Sutton, R & Barto, A. (1981) *Psychol. Rev.* **88**, 135–170.
60. Gross, C. G. (1998) *Brain Vision and Memory: Tales in the History of Neuroscience*. (MIT Press).
61. Ungerleider, L & Mishkin, M. (1982) *In Analysis of Visual Behavior*. (MIT Press).

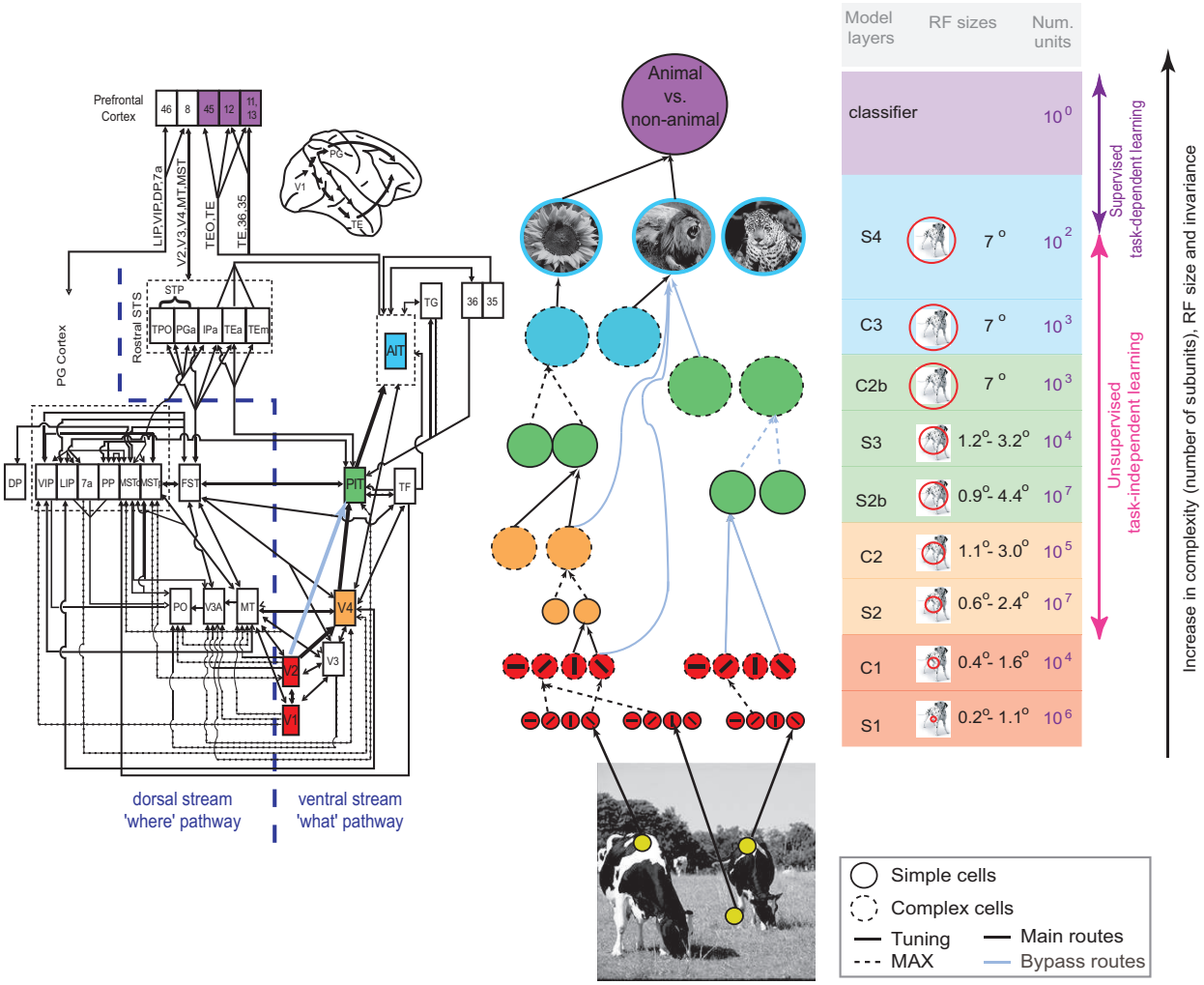

Fig. 1. Sketch of the model. Tentative mapping between the ventral stream in the primate visual system (left) and the functional primitives of the feedforward model (right). The model accounts for a set of basic facts about the cortical mechanisms of recognition that have been established over the last decades: From V1 to IT, there is an increase in invariance to position and scale [1, 2, 4, 5, 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 visual cortex, is to achieve a trade-off between selectivity and invariance within a hierarchical architecture. As in [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 in the figure). The tuning of the S_2, S_{2b} and S_3 units (corresponding to V2, V4 and PIT) is determined here by a prior developmental-like unsupervised learning stage (see *SI*). 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 while the bypass route [40] is denoted with blue arrows (see *SI*). 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 [60] (with permission by the author) which represents a juxtaposition of the diagrams of [61, 48].

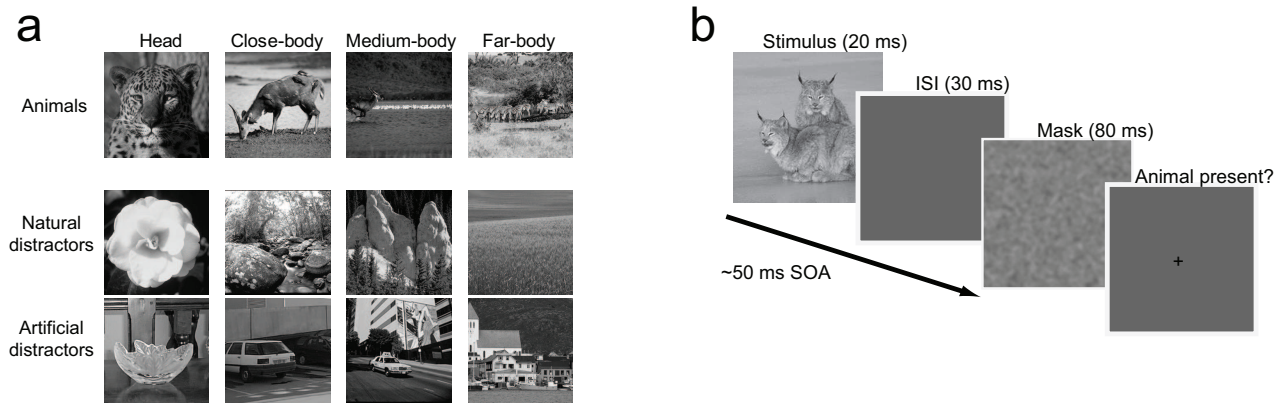

Fig. 2. Animal vs. non-animal categorization task. a) The four (balanced) classes of stimuli. Animal images (a subset of the image database used in [32]) were manually arranged into four groups (150 images each) based on the distance of the animal from the camera: head (close-up), close-body (animal 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*). 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.

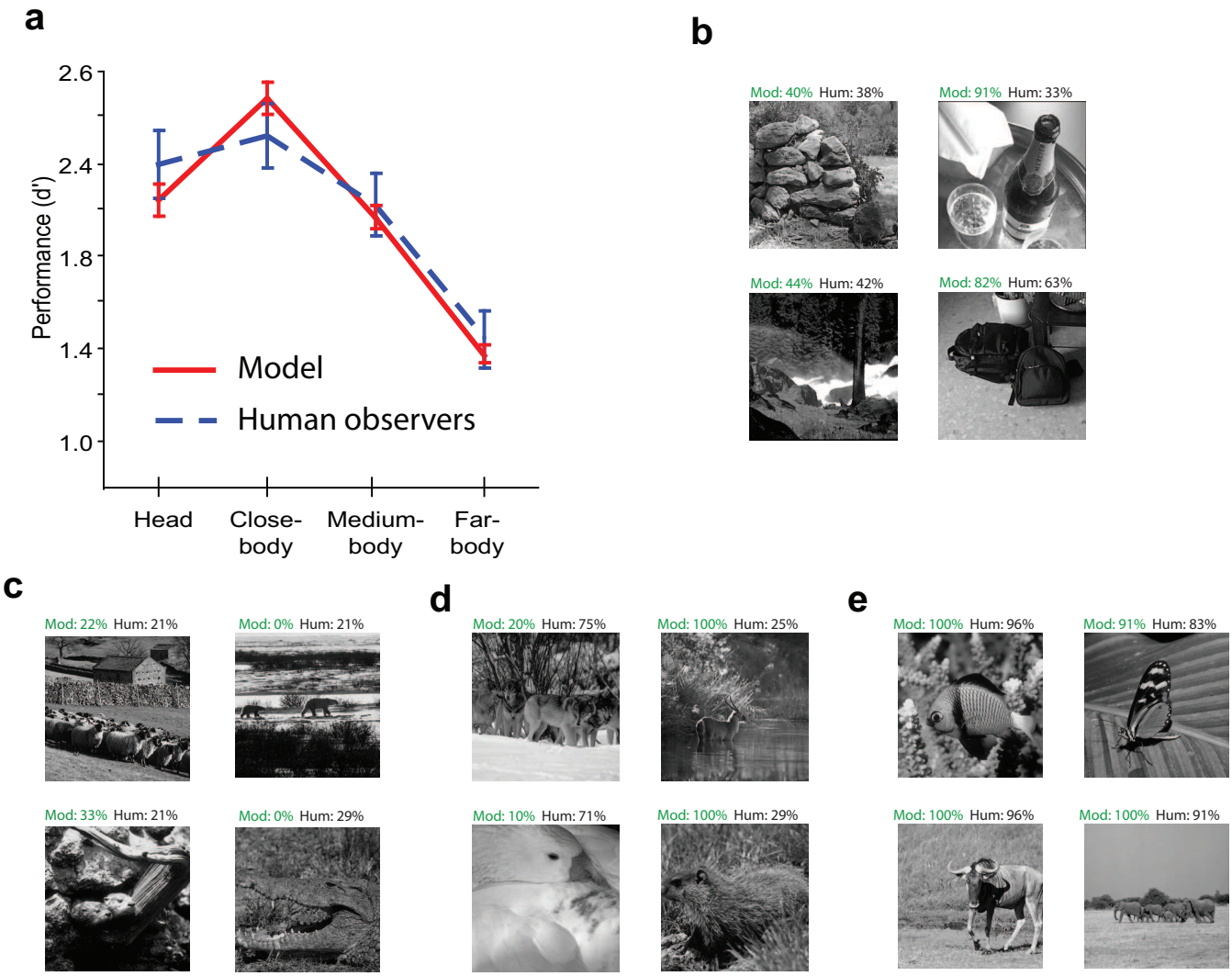

Fig. 3. Comparison between the model and human observers. a) Model vs. human-level accuracy. Human observers and the model exhibit a very similar pattern of performance (measured with d' measure, see *SI*). Error bars indicate the standard errors for the model (computed over $n = 20$ random runs) and 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).