The limits of feed-forward processing in visual object recognition

Gabriel Kreiman[1,2], Thomas Serre[3] and Tomaso Poggio[3]

[1] Ophthalmology and Neuroscience, Children's Hospital Boston, Harvard Medical School
[2] Center for Brain Science, Harvard University
[3] McGovern Institute for Brain Research, Massachusetts Institute of Technology

1. Motivation

Visual recognition can be very fast

- Psychophysics studies show fast recognition (Potter and Levy 1969)
- Object recognition can occur in the near absence of attention (Li et al 2002)

Selective physiological signals show very short latencies

- Scalp EEG signals suggest categorization in ~150 ms (Thorpe et al 1996)
- Single unit studies show selectivity in ~ 100 ms (e.g. Hung et al 2005)

The visual architecture includes forward and back-projections

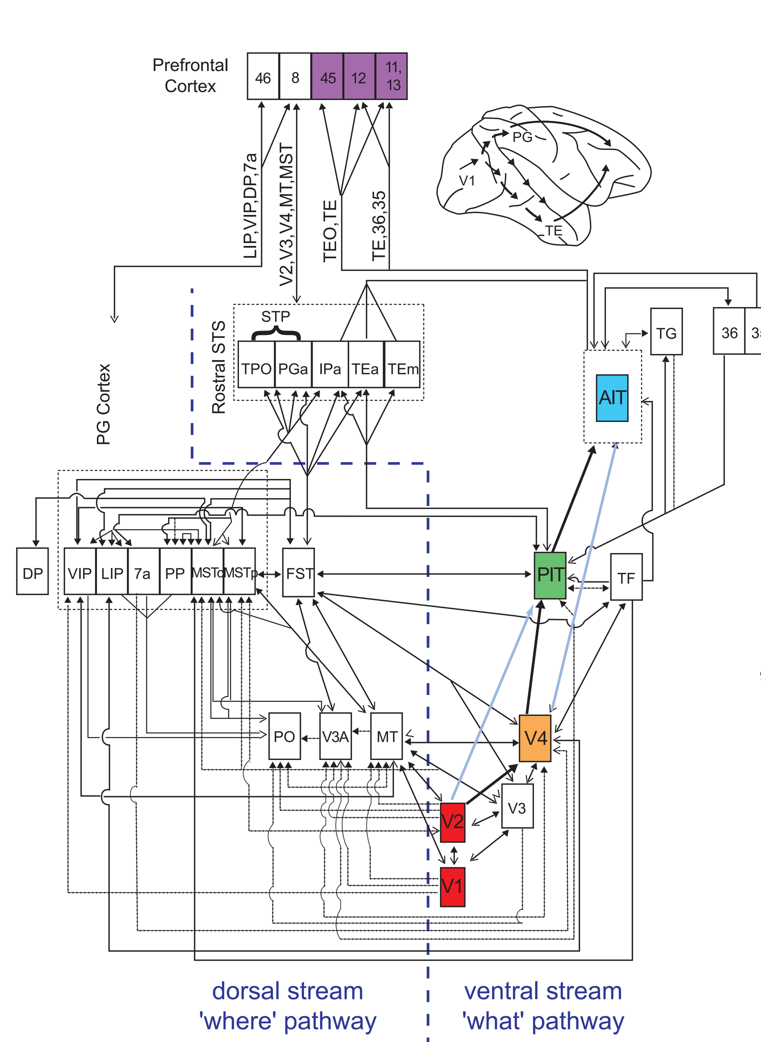
- There are massive back-projections in the visual system (Felleman et al 1991)

Yet some tasks appear to be more difficult

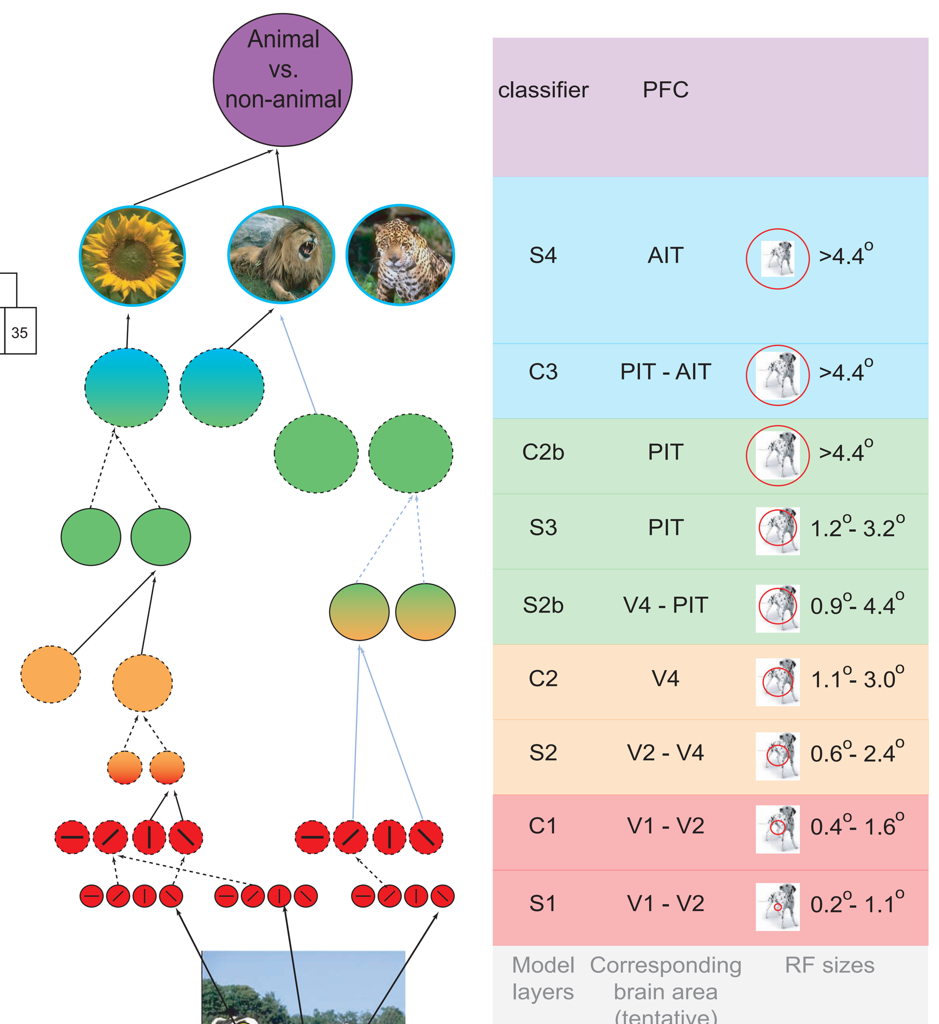
- Visual search may take several hundred ms (Wolfe et al 2004)

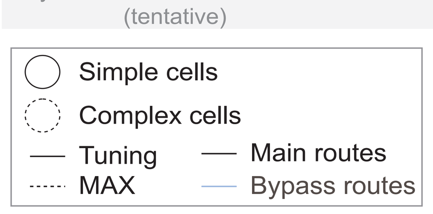
2. Methods

2.1 A hierarchical feed-forward object recognition model



(Felleman and Van Essen, 1991)





Extending Hubel and Wiesel's ideas; Riesenhuber and Poggio 1999, Serre et al 2005; see also Wallis et al 1997, Mel 1997, Fukushima 1980, Perrett et al 1992, LeCun et al 1998

Tuning operation

$$y = g\left(\frac{\sum_{j=1}^{n} w_j x_j^{p}}{k + \left(\sum_{j=1}^{n} x_j^{q}\right)^{r}}\right)$$

Soft-max operation

$$y = g\left(\frac{\sum_{j=1}^{n} x_j^{q+1}}{k + \left(\sum_{j=1}^{n} x_j^{q}\right)}\right)$$

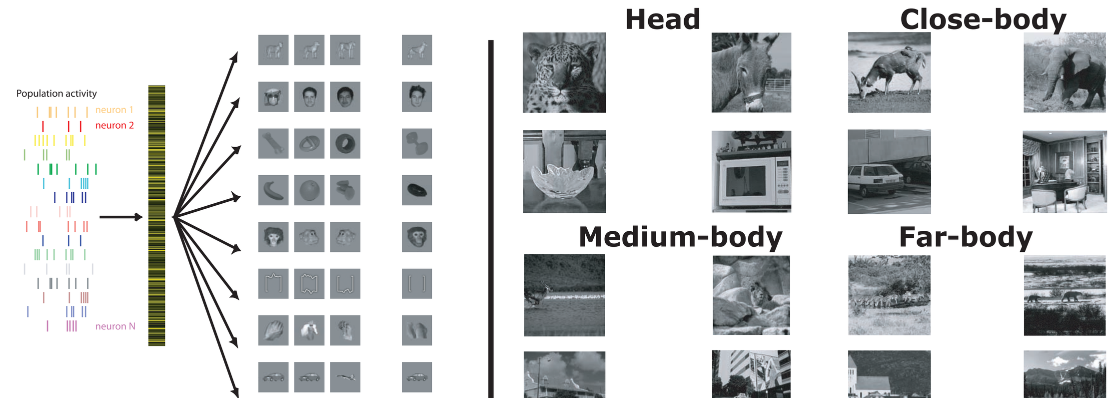
2.2 Supervised learning

Support Vector Machine with linear kernel used for classification

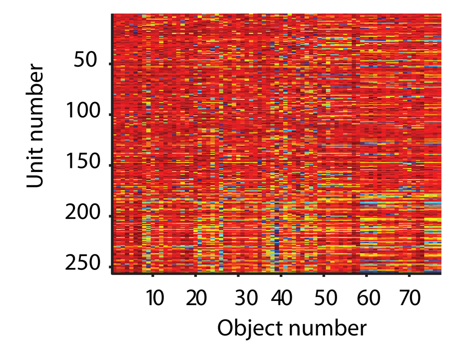
- g classes, (e.g. G_1, ..., G_{77})
- for each G_i, build binary classifier $f(x) = \sum_i c_i k(x, x_i)$, linear kernel, $f(x) = \sum_i^n w_i x_i$
- separate training and testing sets
- one-versus-all classification: take prediction that maximizes classifier output

2.3 Images

Training with model responses to 77 grayscale isolated objects unless otherwise stated.

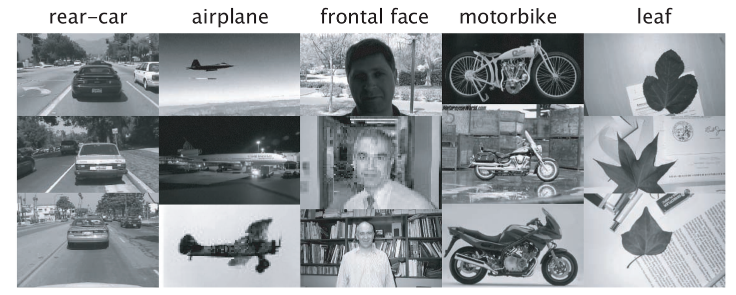


3. The model can explain many physiological observations



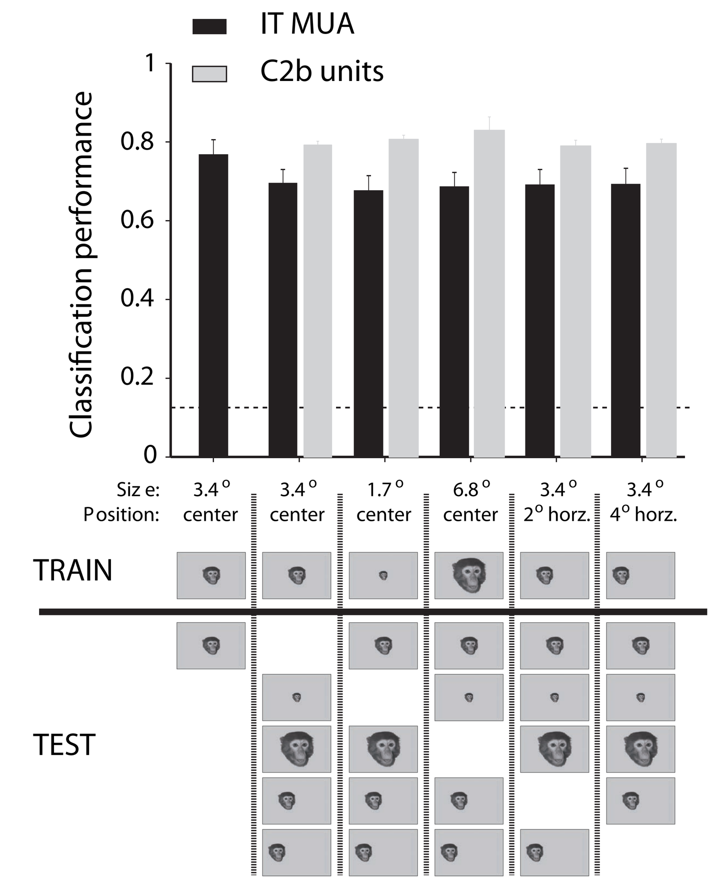
Sample model responses from 256 randomly selected C2b units to 77 grayscale isolated objects

Datasets			AI systems	Model
(CalTech)	Leaves	[Weber et al., 2000b]	84.0	97.0
(CalTech)	Cars	[Fergus et al., 2003]	84.8	99.7
(CalTech)	Faces	[Fergus et al., 2003]	96.4	98.2
(CalTech)	Airplanes	[Fergus et al., 2003]	94.0	96.7
(CalTech)	Motorcycles	[Fergus et al., 2003]	95.0	98.0
(MIT-CBCL)	Faces	[Heisele et al., 2002]	90.4	95.9
(MIT-CBCL)	Cars	[Leung, 2004]	75.4	95.1



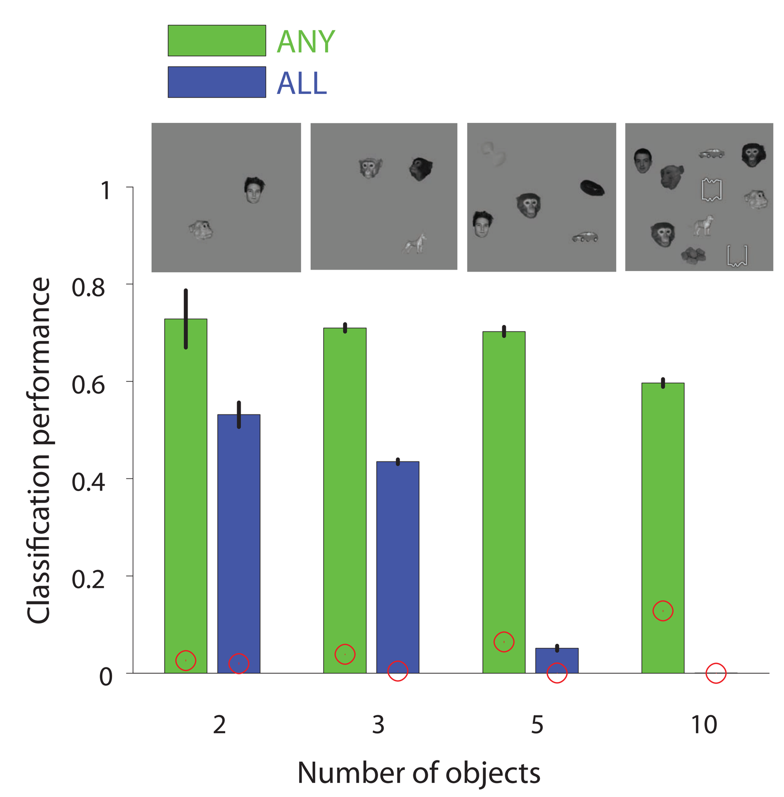
Serre, Wolf, Poggio 2005

The model can perform quite well in multiple standard data sets.



Classification based on model C2b units generalizes over changes in scale and position which is similar to the pattern of generalization seen in the readout from populations of neurons in inferior temporal cortex

4. Performance drops with increasing number of objects in the image

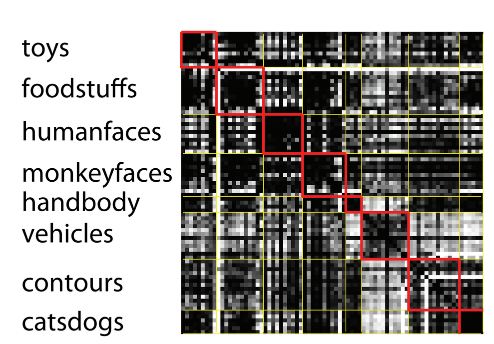


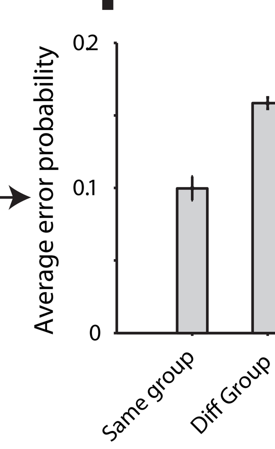
SVM classifier trained on 256 C2b responses to 77 isolated objects

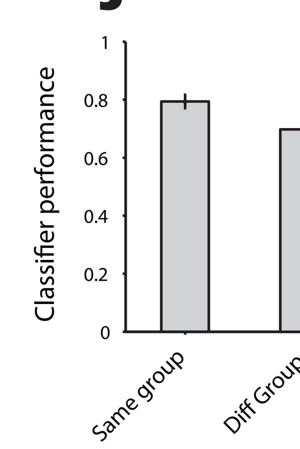
Performance tested on images containing multiple objects (either 2, 3, 5 or 10 objects) at random positions (with no overlap).

Two possible questions:
-"ANY": hit = single classifier prediction matches *any* objects present in the image
-"ALL": hit = multiple classifier predictions match *all* objects present in the image

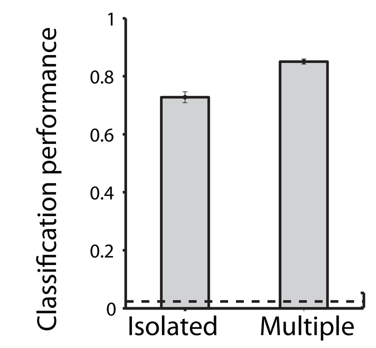
5. Interference depends on object similarity







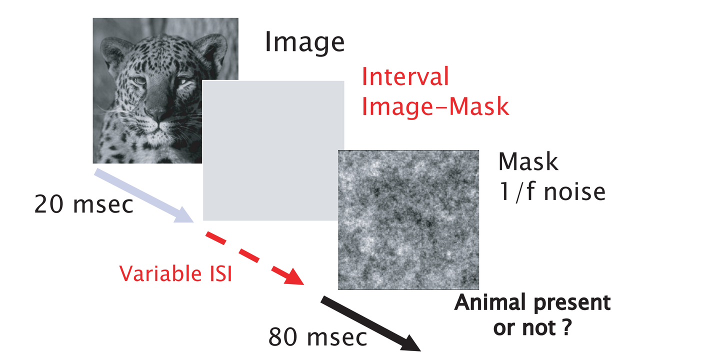
6. Read-out performance increases when training in clutter

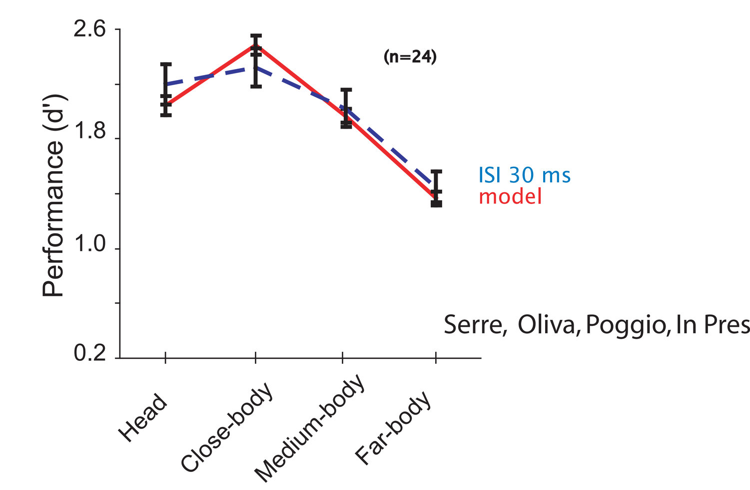


SVM classifier trained with responses of 256 C2b units to either *Isolated* objects or images with two (*Multiple*) objects

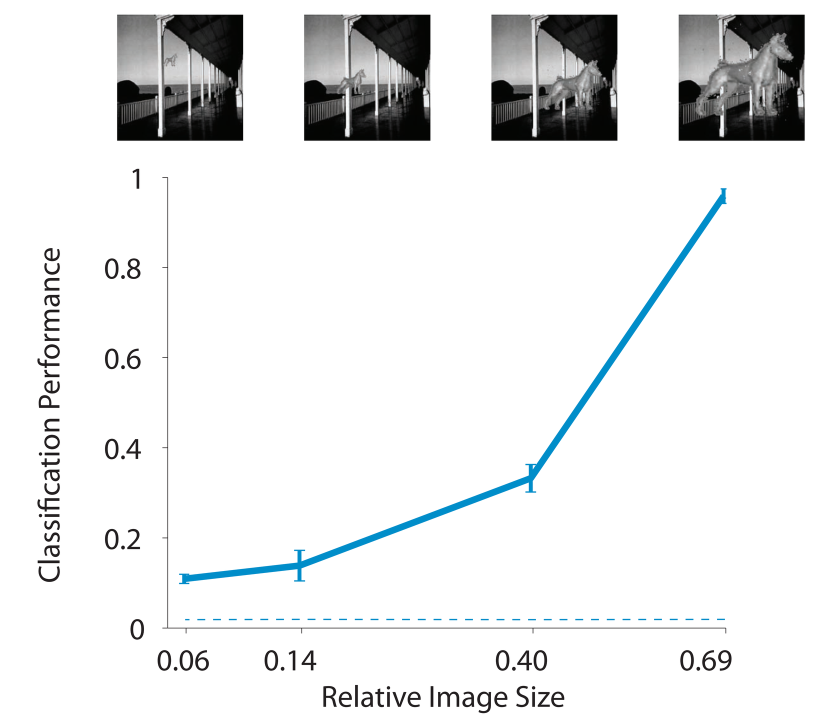
Performance tested on images with 2 objects

7. Performance drops when target is small compared to background





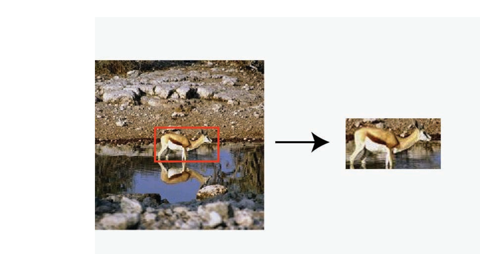
The model performance on a complex categorization task is comparable to human psychophysical measurements under masking. Performance falls significantly in the "far" condition.



SVM classifier trained on 256 C2b responses to 77 isolated objects

Performance tested on images containing those objects embedded in 100 natural background images.

8. A possible solution



Performance in clutter conditions can improve to the level of human performance with long ISIs (60ms) after cropping relevant parts of the image.

9. Summary

- A feedforward architecture can provide a selective and robust response suitable for immediate object recognition tasks.

- Clutter due to multiple objects and background impairs performance. Object recognition under clutter may require additional mechanisms (e.g. feedback).

10. References

Potter M., Levy, E. (1969). Recognition memory for a rapid sequence of pictures. Journal of Experimental Psychology 81, 10-15.
Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. Nature 381, 520-522.
Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex (Boston, MIT), pp. CBCL Paper #259/AI Memo #2005-2036.
Hung, C., Kreiman, G., Poggio, T., and DiCarlo, J. (2005). Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex. Science 310, 863-866.
Walther, D., Rutishauser, U., Koch, C., and Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Computer Vision and Image Understanding 100, 41-63.
Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1, 1-47.
Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019-1025.
Wolfe, Horowitz (2004). What attributes guide the deployment of visual attention and how do they do it? Nat Rev Neurosci 5, 495-501
LeCun Y, Bottou, Bengio, Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, 86, 2278-2324, Nov. 1998..