Vision: Learning in the Brain

Thomas Serre

Center for Biological and Computational Learning
Brain and Cognitive Sciences Department

Object Recognition by Humans

- Some numbers:
  - ❑ Basic-level categories*** ~ thousands
  - ❑ Subordinate-level:
    - ✓Specific scenes** ~ thousands
    - ✓Faces* > hundreds

*** Biederman (1974)
** Standing (1973): good memory for 10,000 photos
* Bahrick et al. (1975): 90% recognition of year-book photos of schoolmates, indep. of class size (90 to 900), and of time elapsed since graduation (3 mos. - 35 years).

Source: Modified from DiCarlo & Kanwisher (9.916)

M. Potter (1971)

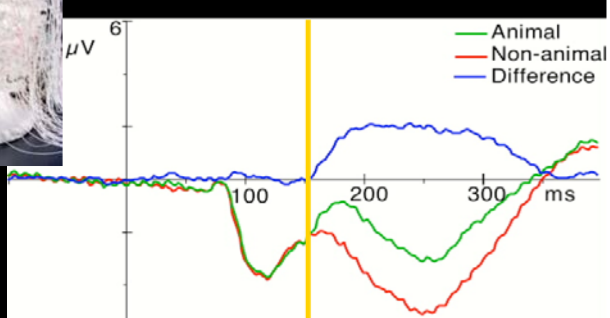
- Measured rate of scene/object recognition using RSVP
- Subjects able to get the “gist” even at 7 images/s from an unpredictable random sequence of natural images
  - ❑ No time for eye movements
  - ❑ No top-down / expectations

Source: Modified from DiCarlo & Kanwisher (9.916)

Thorpe et al. (1996)

- Rapid animal vs. non-animal categorization
  - ❑ Animal present vs. absent
- How fast is object recognition?

ERP study: Recognition under 150 ms



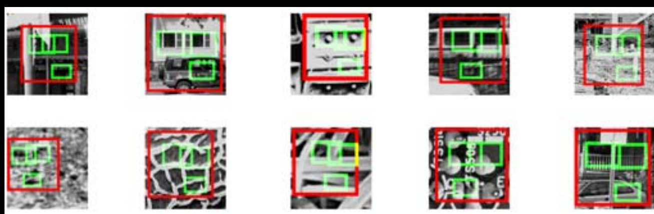
[Thorpe et al, Nature 1996]

Robustness to degradations

[Yip and Sinha, 2002]

- Recognition performance:
  - ❑ 7 x 10 pix: more than ½ set of familiar faces
  - ❑ 19 x 27 pix: ceiling level
- Typical computer vision algorithms:
  - ❑ 60 x 60 pixels [Heisele et al., 2001, 2002, 2004]
  - ❑ Typical face db > 200 x 200 pix

FAs by one computer vision system



[Heisele, Serre & Poggio, 2001]

- AI systems << primate visual system
- AI systems ~ birds and insects

Pigeons can discriminate between:

- Paintings
  - Monet vs. Picasso
  - Van Gogh vs. Chagall
- Animal vs. non-animal
- Different kind of leaves
- Letters of the alphabet
- Human artifacts vs. natural objects

Source: Modified from Pawan Sinha (9.670)

Bees can discriminate between flower patterns

Training: One pattern reinforced with sugar

Test: New patterns

Source: Modified from Pawan Sinha (9.670)

Understanding how brains recognize objects may lead to better computer vision systems

Roadmap

1. Neuroscience 101:
   - Neuron basics
   - Primate visual cortex

2. Computational model developed at CBCL

Neuron Basics

Different shapes and sizes but common structure

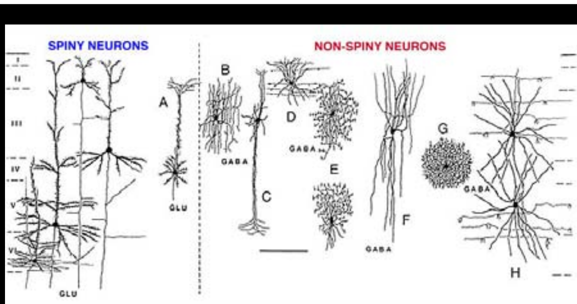
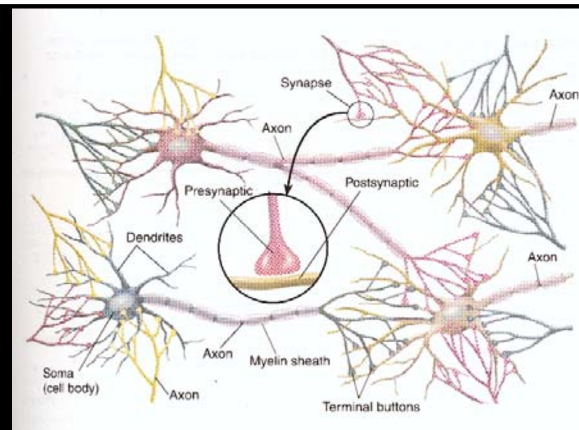


Figure 12. Basic cell types in the monkey cerebral cortex. Left: spiny neurons that include pyramidal cells and stellate cells (A). Spiny neurons utilize the neurotransmitter glutamate (Glu). Right: smooth cells that use the neurotransmitter GABA. B, cell with local axon arcades; C, double bouquet cell; D, H, basket cells; E, chandelier cells; F, bitufted, usually peptide-containing cell; G, neurogliaform cell.

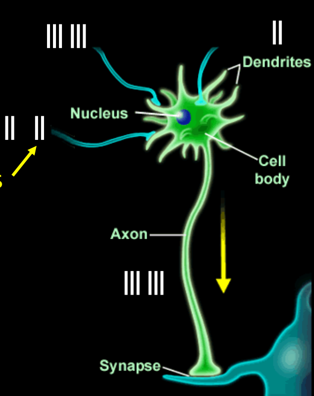
Source: http://webvision.med.utah.edu/

Neural Network



Source: Modified from Jody Culham's web slides

Neuron basics



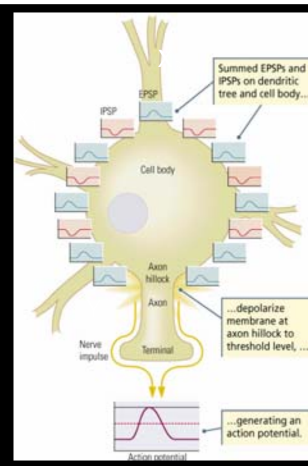
spikes

INPUT
= Digital

COMPUTATION
= Analog

OUTPUT
= Digital

Computation at the SOMA

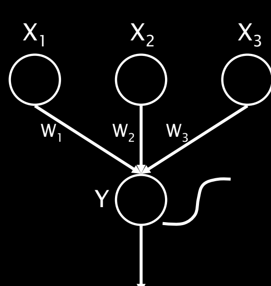


Model Neuron

$$Y = H\left(\sum w_i x_i\right)$$

$$Y = H\left(\sum w_i x_i \Big/ \sqrt{\sum x_i^2}\right)$$

$$Y = \exp\left(-\sum |x_i - w_i|^2 \Big/ \sigma^2\right)$$



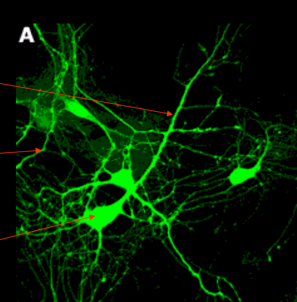
In the real world

- 10^{10}–10^{12} neurons (10^5 neurons/mm^3)
- 10^3 – 10^4 synapses/neuron
- 10^{15} synapses

Axon ~ 5µm

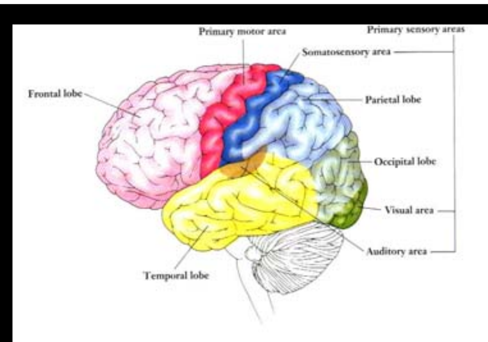
Dendrites ~ 1µm

Soma ~ 20µm



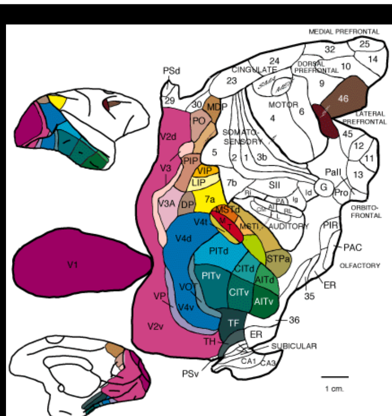
The visual system

Gross Brain Anatomy



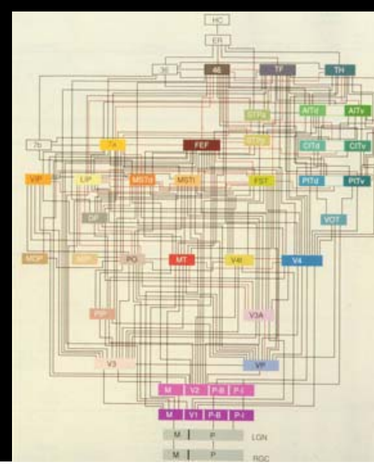
50–60% of the brain devoted to vision

The Visual System



[Van Essen & Anderson, 1990]

The Visual System

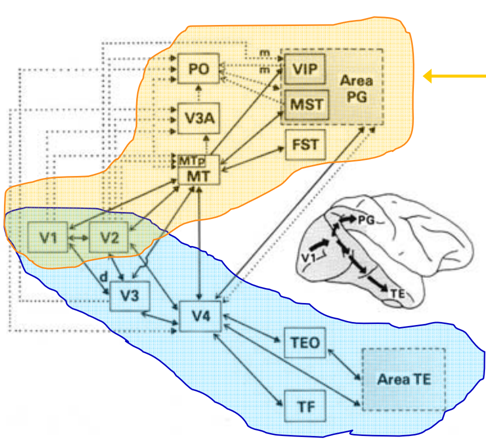


The complexity of the hardware matches that of the problem…

Computational approaches may help (cf. bioinformatics)

[Felleman & van Essen, 1991]

The ventral visual stream



dorsal stream: "where"

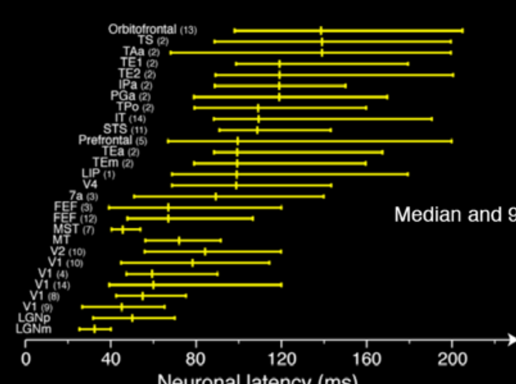
ventral stream: "what"

[Ungerleider & Haxby, 1994]

Object representation in the ventral stream

Timing along the visual pathways consistent with anatomy

Observation: ~10 ms latency difference between each cortical area

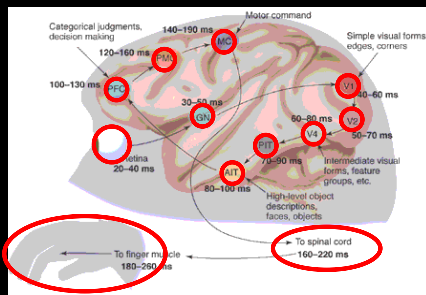


Median and 90% range

Source: Modified from DiCarlo & Kanwisher (9.916)

Adapted from: Nowak and Bullier (199

Feedforward architecture



[Thorpe and Fabre-Thorpe, 2001]

The Retina

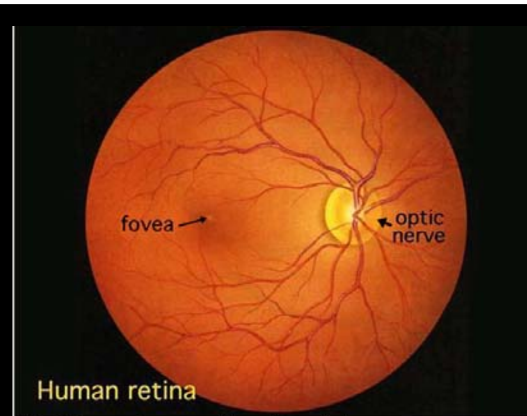
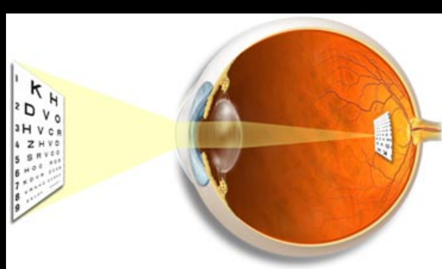


Fig. 1. Human retina as seen through an opthalmoscope.

Source: http://webvision.med.utah.edu/

Photoreceptors

- Back of the eyes
- Behind blood vessels
- 100 million+ photoreceptors



Source: http://webvision.med.utah.edu/

Rods and cones

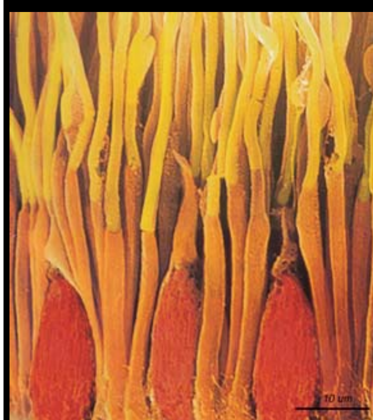


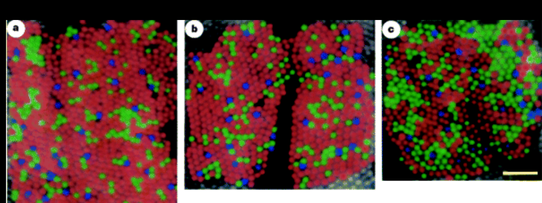
Fig1b. Scanning electron micrograph of the rods and cones of the primate retina. Image adapted from one by Ralph C. Eagle/Photo Researchers, Inc.

- Duplicity theory
- 2 classes of photoreceptors for 2 different luminance regimes:
  - Scotopic vision: Rods
  - Photopic vision: Cones

Source: http://webvision.med.utah.edu/

Cone type distrib. varies between ind.

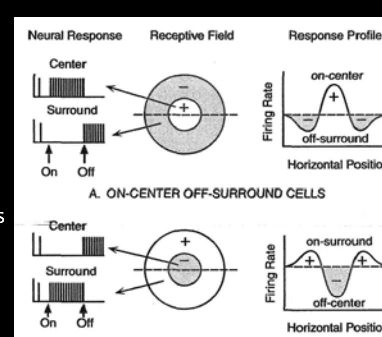
- We don't all see the same thing!!
- Human trichromatic cone mosaic



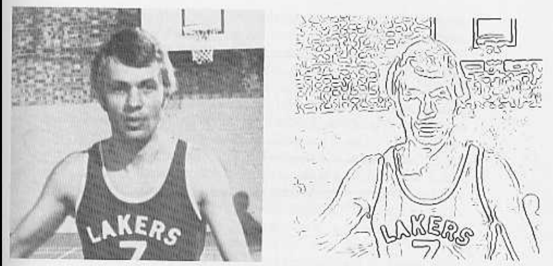
[Roorda & Williams, Nature 1999]

Ganglion cells / LGN

- Center / surround receptive fields
- Convolution / Laplacian:
  - Enhances local changes / boundaries
  - Disregard smooth surfaces
  - Computation of zero-crossings



Ganglion cells / LGN

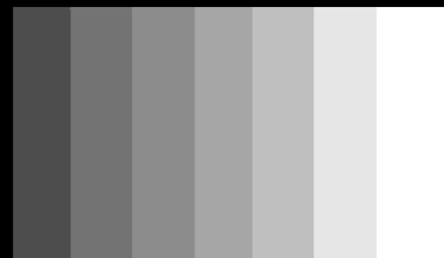


(a) (b)

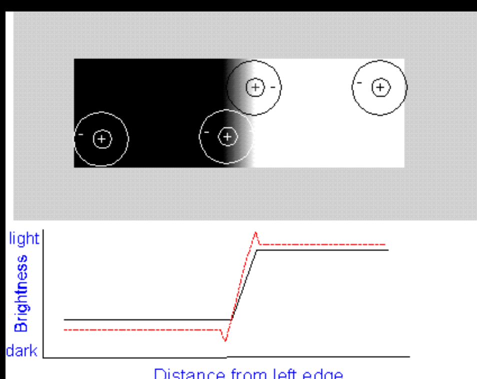
Figure 2–15. Another example of zero-crossings; here, the intensity of the lines has been made to vary with the slope of the zero-crossing, so that it is easier to see which lines correspond to the greater contrast. (Courtesy BBC Horizon.)

[Marr, Vision 1982]

Illusions: Mach Bands



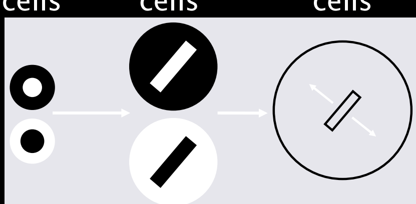
Illusions: Mach Bands

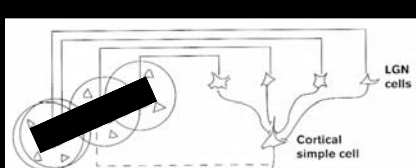


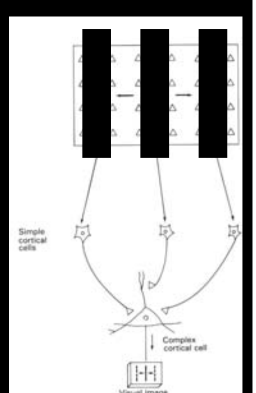
V1: Orientation selectivity

V1: hierarchical model

LGN-type cells | Simple cells | Complex cells

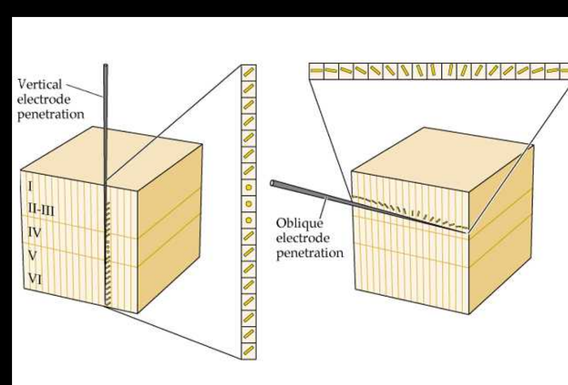




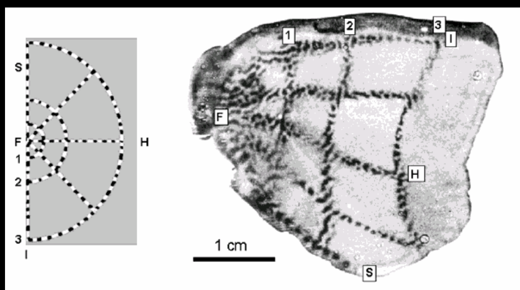


(Hubel & Wiesel, 1959)

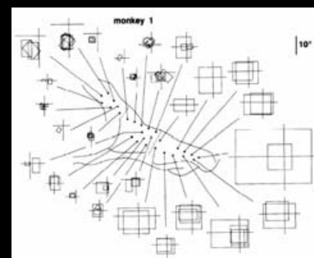
V1: Orientation columns

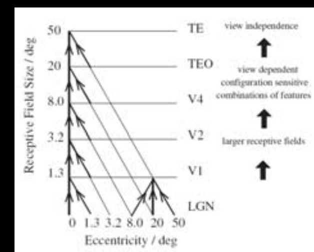


V1: Retinotopy

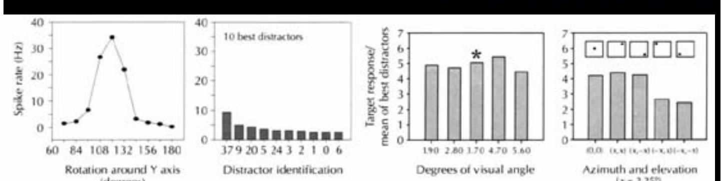


Beyond V1: A gradual increase in RF size

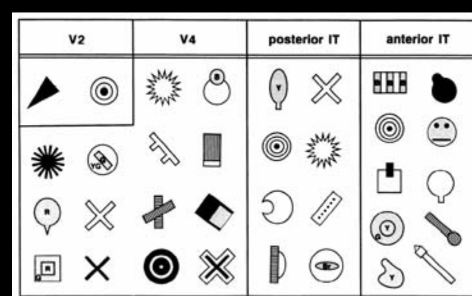




Beyond V1: A gradual increase in invariance to translation and size



Beyond V1: A gradual increase in the complexity of the preferred stimulus

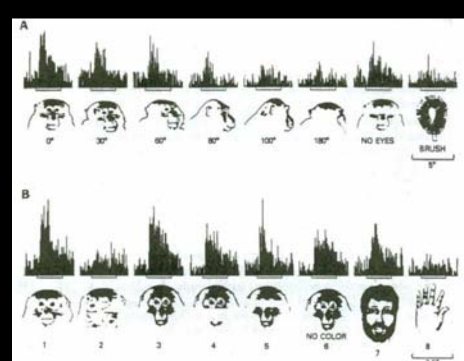


[Kobatake et al, 1994]

Anterior IT

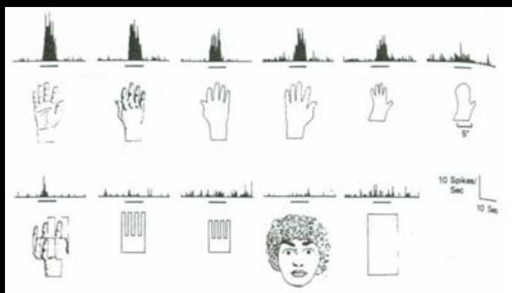
- Very large receptive fields (several degrees)
- Invariance:
  - Position
  - Scale
- Hand, face, "toilet brush" cells, etc
- Broad cells tuning
  - Population coding
  - ≠"grand-mother" cells

AIT: Face cells



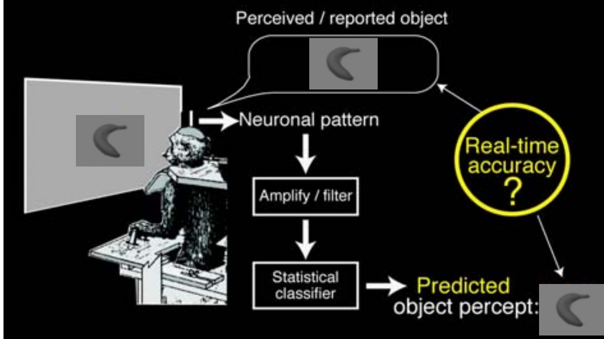
[Desimone et al. 1984]

AIT: Hand cells



[Desimone et al. 1984]

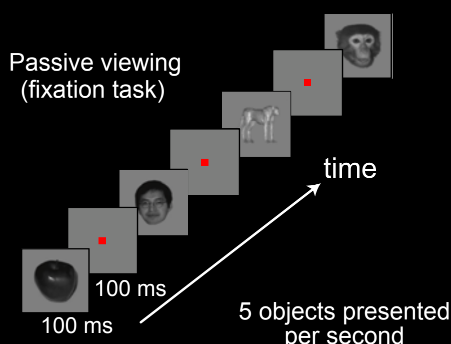
Is it possible to read out what the monkey is seeing?



[Hung, Kreiman, Poggio & DiCarlo, 2005]

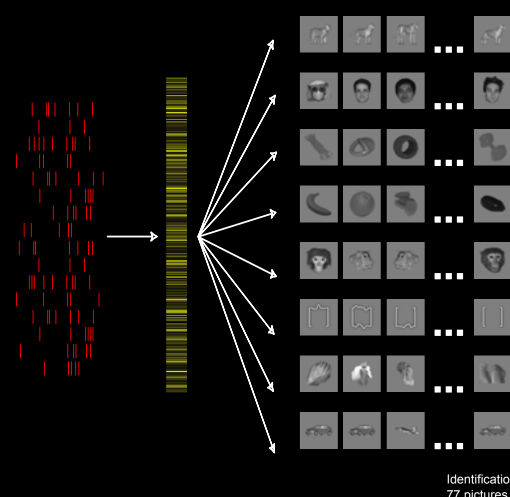
Source: Courtesy of Gabriel Kreiman

Stimulus presentation



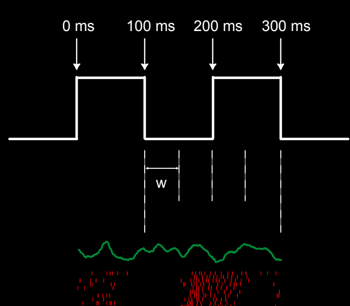
- 10-20 repetitions per stimulus
- presentation order randomized
- 77 stimuli drawn from 8 pre-defined groups

Source: Courtesy of Gabriel Kreiman



[Hung, Kreiman, Poggio & DiCarlo, 2005]

Input to the classifier



SUA: spike counts in each bin

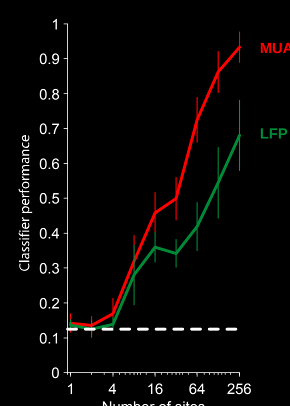
MUA: spike counts in each bin

LFP: power in each bin

MUA+LFP: concatenation of MUA and LFP

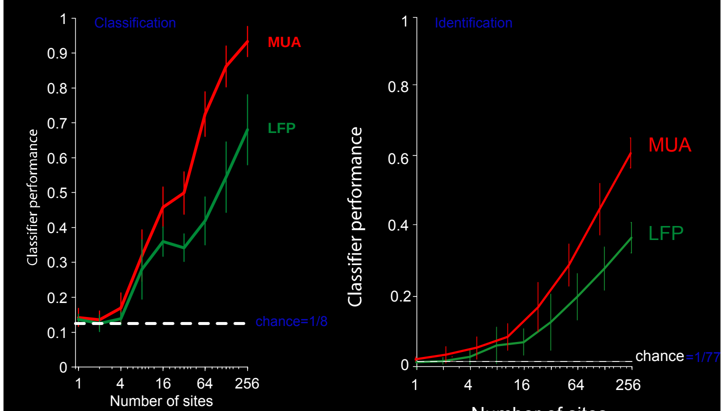
Source: Courtesy of Gabriel Kreiman

Object category can be decoded quite accurately from the population response



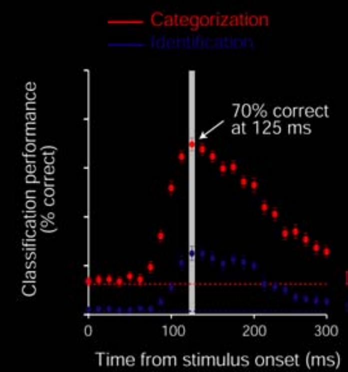
[Hung, Kreiman, Poggio & DiCarlo, 2005]

Object identity



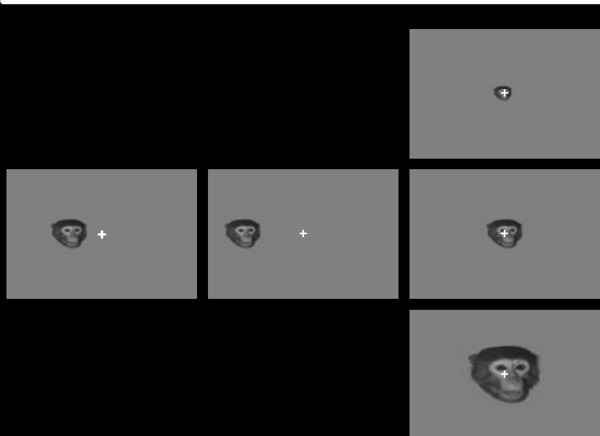
[Hung, Kreiman, Poggio & DiCarlo, 2005]

12.5 ms are enough to decode well above chance



[Hung, Kreiman, Poggio & DiCarlo, 2005]

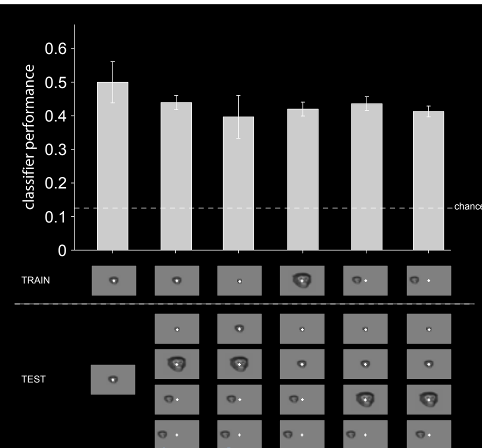
Scale and translation invariance



[Hung, Kreiman, Poggio & DiCarlo, 2005]

Source: Courtesy of Gabriel Kreiman

The classifier extrapolates to new scales and positions

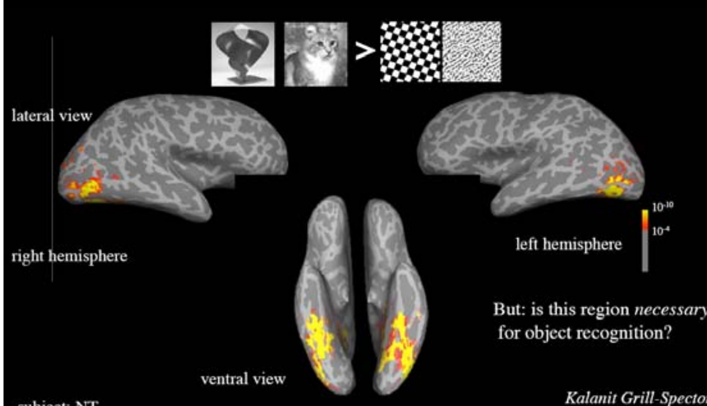


[Hung, Kreiman, Poggio & DiCarlo, 2005]

Source: Courtesy of Gabriel Kreiman

What about humans?

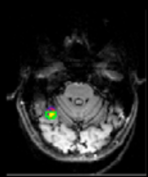
Object selective region in the human brain: LOC

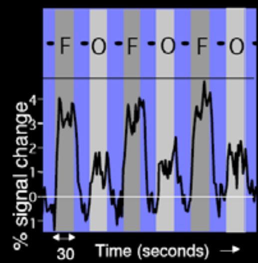


Source: Modified from DiCarlo & Kanwisher (9.916)

Face selective region in the human brain: FFA

Faces > Objects





[Kanwisher, McDermott & Chun, 1997]

Source: Modified from DiCarlo & Kanwisher (9.916)

Single neuron recordings in epileptic patients

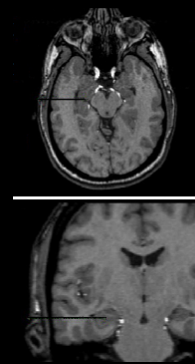
MRI

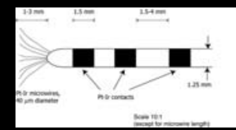
Electrodes

•Patients with pharmacologically intractable epilepsy

•Multiple electrodes implanted to localize seizure focus

•Targets include the hippocampus, entorhinal cortex, amygdala and parahippocampal gyrus



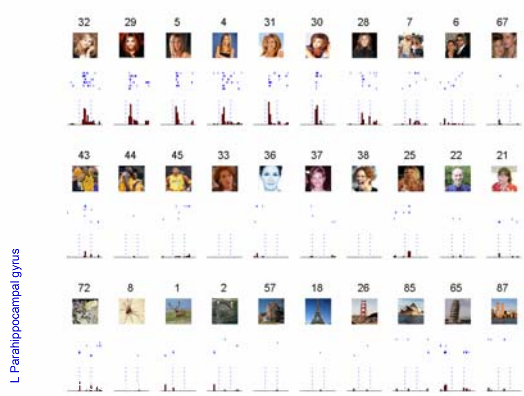


Fried et al *J. Neurosurgery* 1999

Kreiman, Koch, Fried (2000) *Nature Neuroscience* **3**:946-953

Source: Courtesy of Gabriel Kreiman

A neuron selective to actress Jennifer Aniston



Quian Quiroga, Reddy, Kreiman, Koch, Fried. *Nature 2005*
Quiroga, Reddy, Kreiman, Koch, Fried. *Nature, 2005*

137/998 (14%) selective units, 52/137 (38%) showed invariance

Source: Courtesy of Gabriel Kreiman

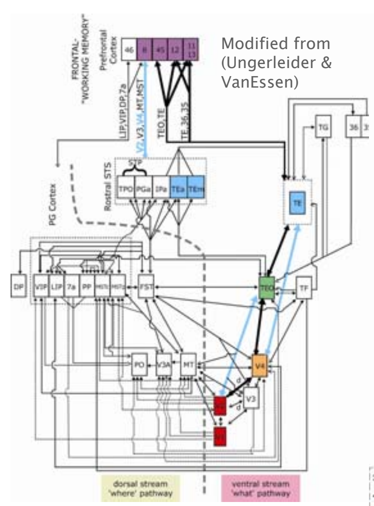
Computational models and work at CBCL

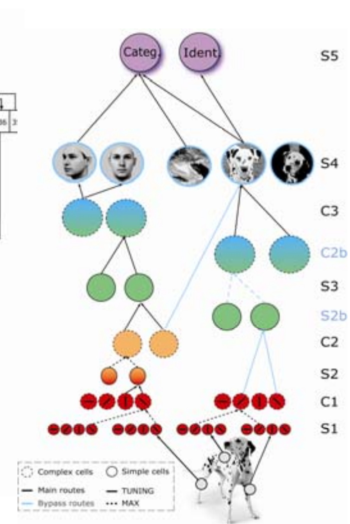
Roadmap

I. The model

II. Comparison with other computer vision systems

III. Comparison with human observers

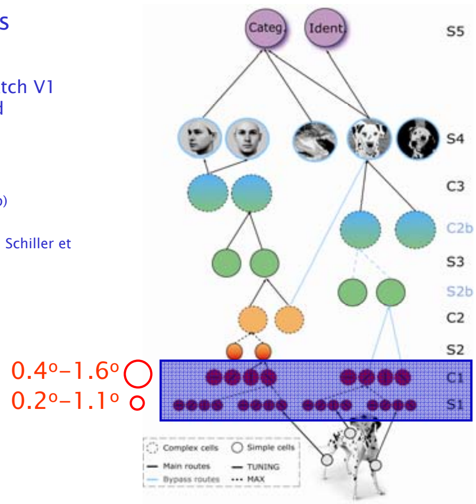




[Riesenhuber & Poggio, 1999, 2000;
Serre, Kouh, Cadieu, Knoblich, Kreiman & Poggio, 2005]

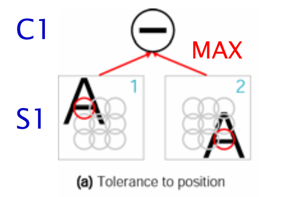
S1 and C1 units

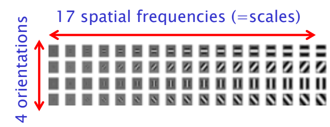
- Tuning properties match V1 parafoveal simple and complex cells
- Assessed with:
  - Gratings
    (Devalois et al, 1982a,b)
  - Bars and edges
    (Hubel & Wiesel, 1965, Schiller et al, 1976a,b,c)

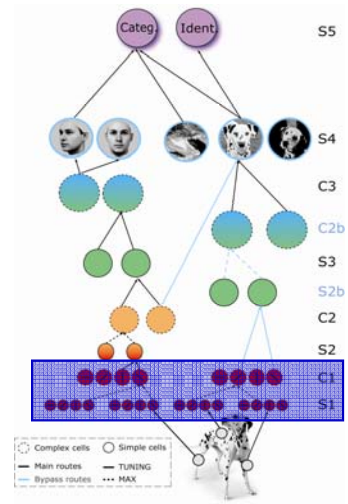


S1 and C1 units

- Increase in tolerance to position (and in RF size)

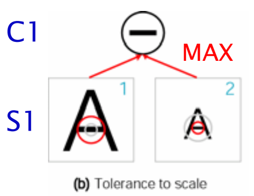


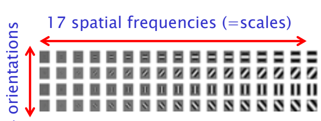


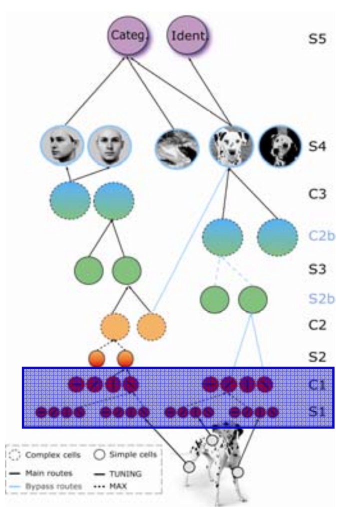


S1 and C1 units

- Increase in tolerance to scale (broadening in frequency bandwidth)

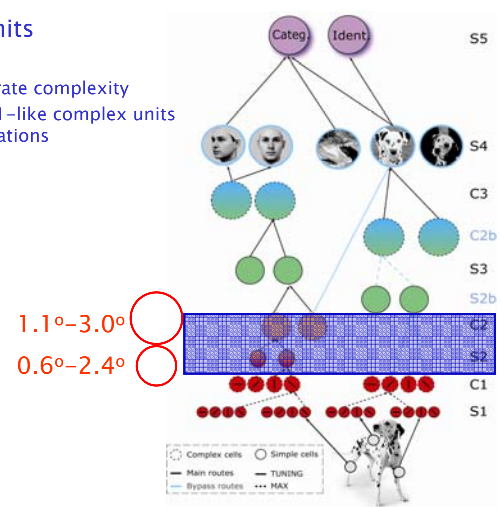






S2 and C2 units

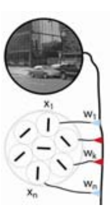
- Features of moderate complexity
- Combination of V1-like complex units at different orientations

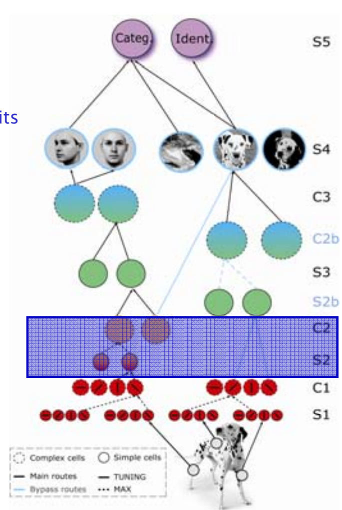


S2 and C2 units

- Features of moderate complexity
- Combination of V1-like complex units at different orientations
  - 10 subunits
  - Synaptic weights **w** learned from natural images

S2 unit





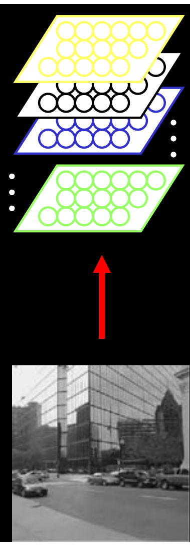
Learning the tuning of units in the model

- Learning is likely to play a key role in the recognition ability of primates
- From V2 to IT in the model, units are tuned to a large number of "patches" from natural images
- Details still open-ended (more than the rest of the model, i.e., RF sizes, tuning properties) for which we have quantitative data
- For clarity, I will describe the learning approach in a more "algorithmic" way (but see thesis for more biological implementation)

Start with S2 layer

Units are organized in n feature maps

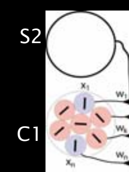
Database ~1,000 natural images



Start with S2 layer

Pick 1 unit from the first map at random

Store in unit synaptic weights the precise pattern of subunits activity, i.e. **w**=**x**



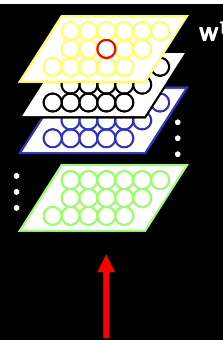


Image "moves" (looming and shifting)

Weight vector **w** is copied to all units in feature map 1 (across positions and scales)

Then pick 1 unit from the second map at random

Store in unit synaptic weights the precise pattern of subunits activity, i.e. **w**=**x**

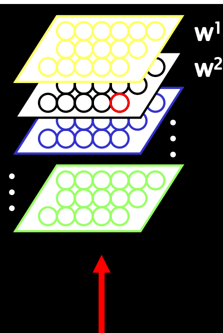


Image "moves" (looming and shifting)

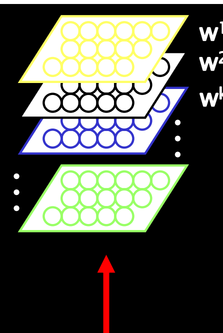
Weight vector **w** is copied to all units in feature map 1 (across positions and scales)

Iterate until k feature maps have been learned

Then present second image

Learn k feature maps

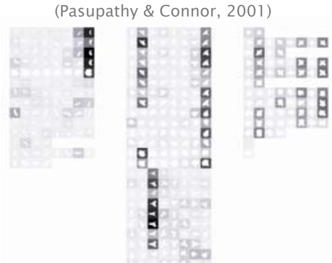
Iterate until all maps have been trained

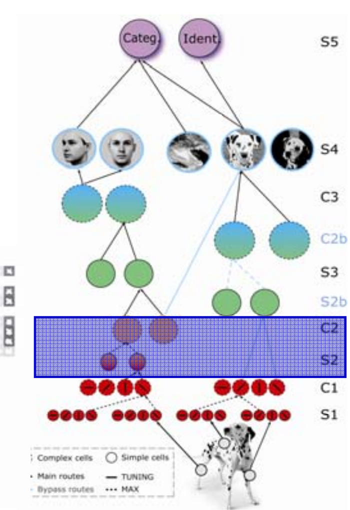


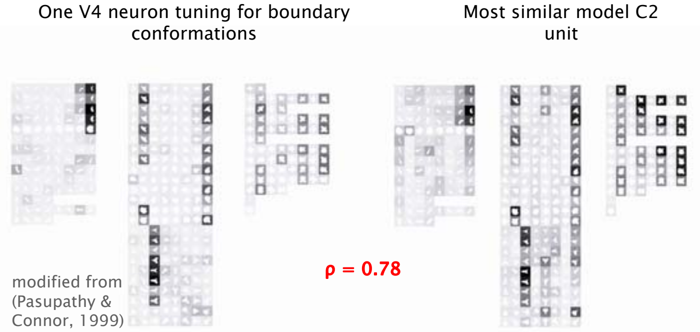
S2 and C2 units

- n=2,000 feature maps total
- Quantitative agreement:
  - Compatible with tuning for boundary conformations in V4

(Pasupathy & Connor, 2001)





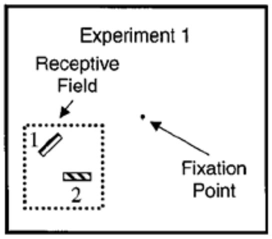


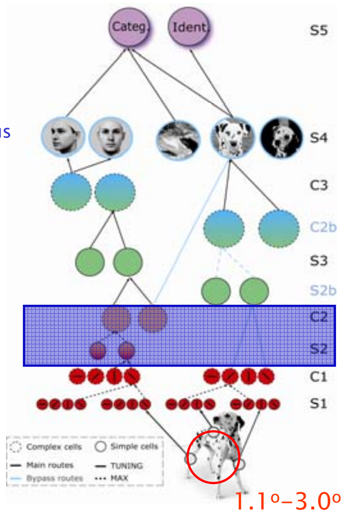
(Serre, Kouh, Cadieu, Knoblich, Kreiman and Poggio, 2005)

S2 and C2 units

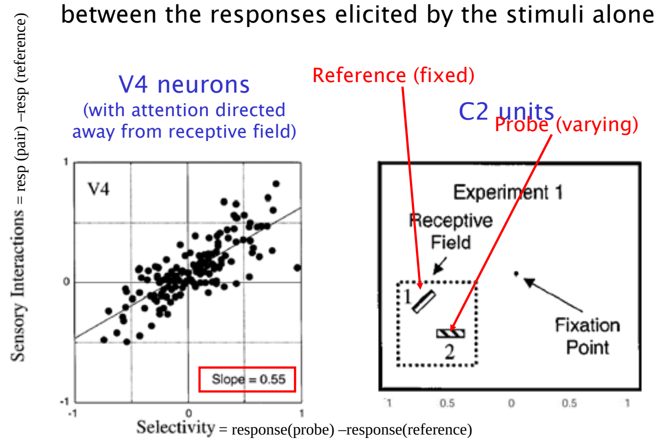
- n=2,000 feature maps
- Quantitative agreement:
  - Compatible with two-bar stimulus presentation

(Reynolds et al, 1999)





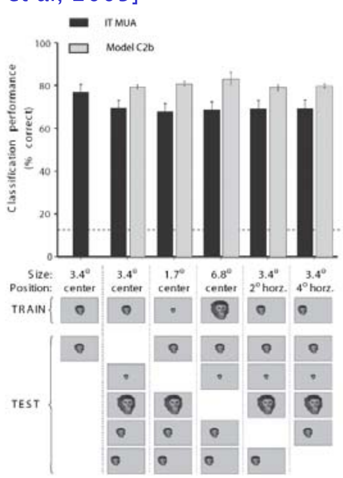
Prediction: Response of the pair is predicted to fall between the responses elicited by the stimuli alone



(Reynolds et al., 1999)

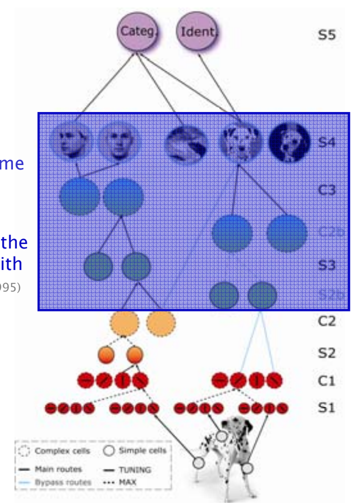
(Serre, Kouh, Cadieu, Knoblich, Kreiman and Poggio, 2005)

Read out data [Hung et al, 2005]



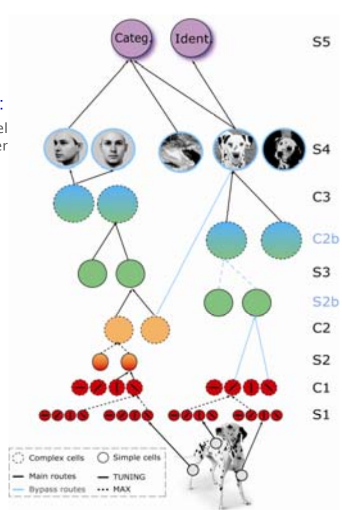
From C2 to S4

- Units are increasingly complex and invariant
- 2,000 "features" at the C3 level ~ same number of feature columns in IT (Fujita et al, 1992)
- Tuning and invariance properties at the S4 level in quantitative agreement with view-tuned units in IT (Logothetis et al, 1995)

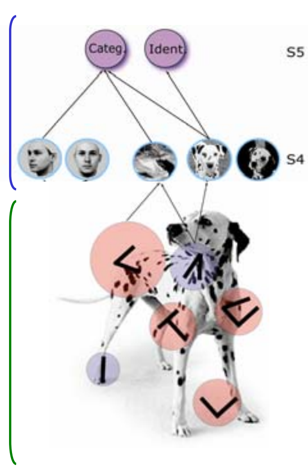


A loose hierarchy

- Bypass routes along with main routes:
  - From V2 to TEO (bypassing V4) (Morel and J.Bullier,1990; Baizer et al., 1991; Distler et al., 1991; Weller and Steele, 1992; Nakamura et al.,1993; Buffalo et al., 2005)
  - From V4 to TE (bypassing TEO) (Desimone et al., 1980; Saleem et al., 1992)
- Some stages are skipped
- Richer dictionary of features with various levels of selectivity and invariance



- Task-specific circuits (from IT to PFC)
  - Supervised learning
  - Linear classifier trained to minimize classification error on the training set (~ RBF net)
  - Generalization capability evaluated on a distinct set of images (test set)
- Features of moderate complexity (from V2 to IT)
  - Unsupervised learning during developmental-like stage
  - From natural images unrelated to any categorization tasks



A neurobiological approach

- Biophysical implementations
  - Based on simple properties of cortical circuits and synapses [Yu et al, 2002; Knoblich & Poggio, 2005]
- Reflects organization of the ventral stream
- Predicts several properties of cortical neurons [Serre, Kouh, Cadieu, Knoblich, Kreiman, Poggio, 2005]

Successful model predictions

- MAX in V1 (Lampl et al, 2004) and V4 (Gawne et al, 2002)
- Differential role of IT and PFC in categ. (Freedman et al, 2001,2002,2003)
- Face inversion effect (Riesenhuber et al, 2004)
- IT read out data (Hung et al, 2005)
- Tuning and invariance properties Of VTUs in AIT (Logothetis et al, 1995)
- Average effect in IT (Zoccolan, Cox & DICarlo, 2005)
- Tow-spot reverse correlation in V1 (Livingstone and Conway, 2003; Serre et al, 2005)
- Tuning for boundary conformation (Pasupathy & Connor, 2001) in V4
- Tuning for Gratings in V4 (Gallant et al, 1996; Serre et al, 2005)
- Tuning for two-bar stimuli in V4 (Reynolds et al, 1999; Serre et al, 2005)
- Tuning to Cartesian and non-Cartesian gratings in V4 (Serre et al, 2005)
- Two-spot interaction in V4 (Freiwald et al, 2005; Cadieu, 2005)

- How well does the model perform on different object categories?
- How does it compare to standard computer vision systems?

Roadmap

I. The model

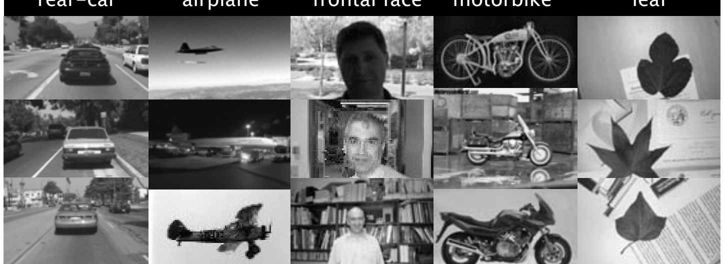
II. Comparison with other computer vision systems

III. Comparison with human observers

CalTech Vision Group

- Constellation models [Leung et al, 1995; Burl et al, 1998; Weber et al., 2000; Fergus et al, 2003; Fei-Fei et al, 2004]

rear-car | airplane | frontal face | motorbike | leaf

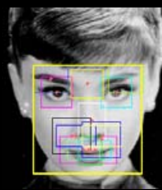


CalTech Vision Group

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

Other approaches

- Hierarchy of SVM-classifiers
  [Heisele, Serre & poggio, 2001, 2002]
  - Component experts
  - Combination classifier



- Fragment-based approach
  [Leung, 2004] based on [Ullman et al, 2002; Torralba et al, 2004]

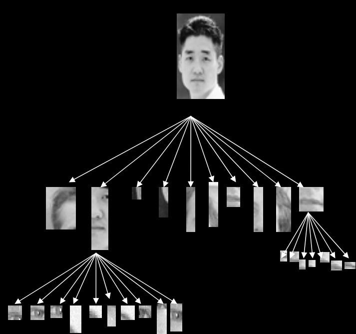
Other approaches

Datasets			AI systems	Model
(MIT-CBCL)	Faces	[Heisele et al., 2002]	90.4	95.9
(MIT-CBCL)	Cars	[Leung, 2004]	75.4	95.1

Near-profile

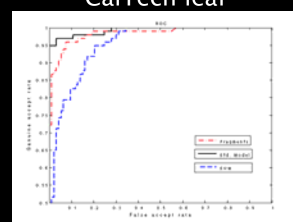
Multi-view car

Fragment-based system

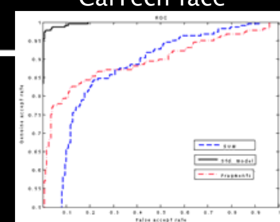


[Ullman et al, 2005; Epshtein & Ullman, 2005]

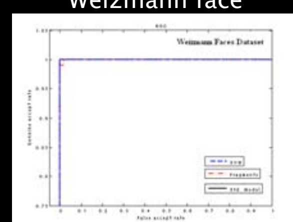
CalTech leaf



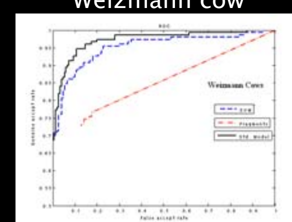
CalTech face



Weizmann face



Weizmann cow



[Chikkerur & Wolf, in prep]; courtesy: Chikkerur

CalTech 101 object dataset

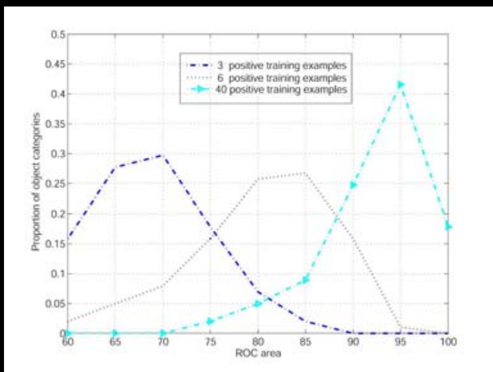
- 40-800 images per categ. (mode ~ 50)
- Large variations in shape, clutter, pose, illumination, size, *etc.*
- Unsegmented (objects in clutter)
- Color information removed

[Fei-Fei et al., 2004]

CalTech 101 object dataset

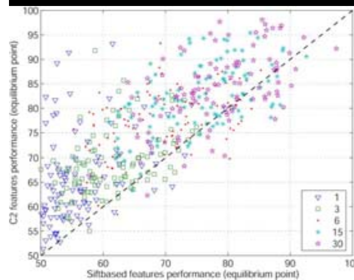
[Fei-Fei et al., 2004]

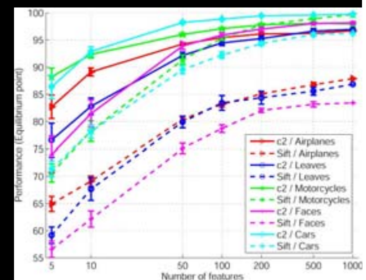
CalTech 101 object dataset



[Serre, Wolf, Poggio, CVPR 2005]

SIFT features [Lowe, 2004]





[Serre, Wolf, Poggio, CVPR 2005]

CalTech 101 object dataset

- Model re-implementation for multi-class
- chance < 1%
- 15 training examples:
  - Serre, Wolf & Poggio (2004) ~ 44%
  - Wolf & Bileschi (in sub) ~ 56%
  - Mutch & Lowe (in sub) ~ 56%
- Others:
  - Holub, Welling & Perona (2005) ~ 44%
  - Berg, Berg & Malik (2005) ~ 45%

StreetScene project

Challenge

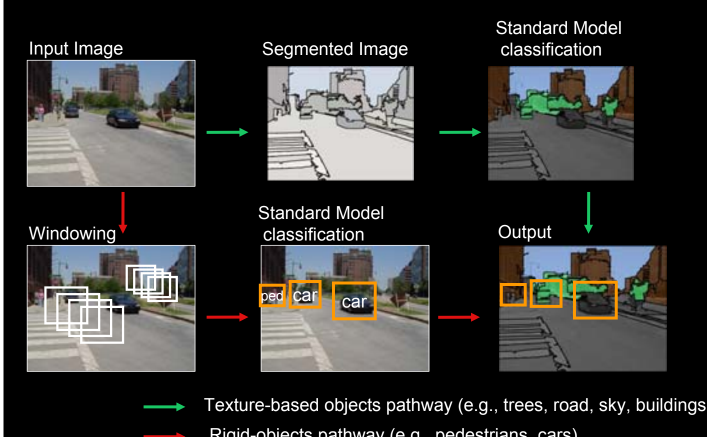
In-class variability:

- Vehicles of different types at many poses, illuminations.
- Trees in both Summer and Winter
- City and suburban scenes

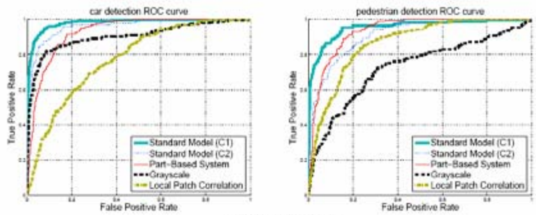
Partial labeling:

- Rigid objects are only labeled if less than 15% occluded.
- Some objects are unlabeled.
- Bounding boxes overlap and contain some background.

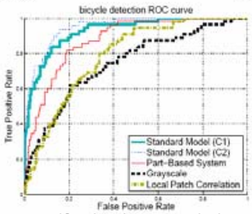
The system



Rigid objects recognition

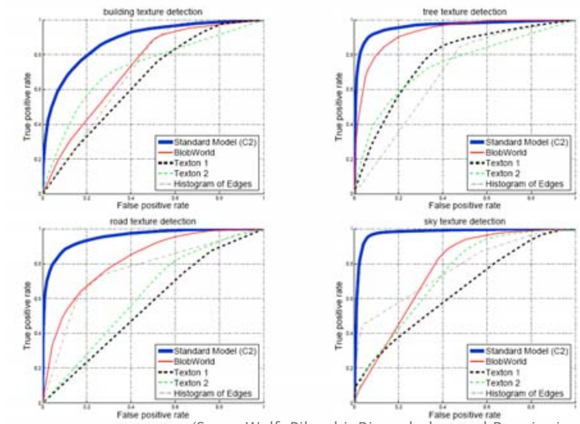


- Local patch correlation: (Torralba et al, 2004)
- Part-based system: (Leibe et al, 2004)



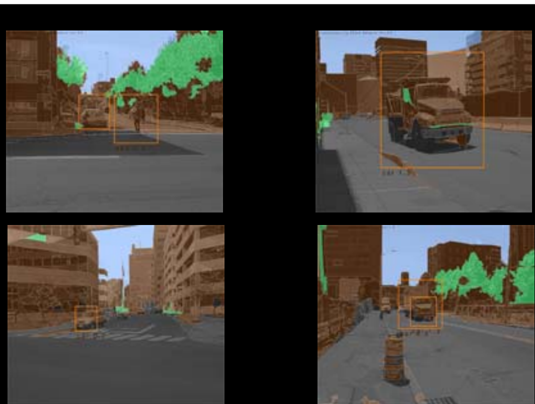
(Serre, Wolf, Bileschi, Riesenhuber and Poggio, in sub)

Textured-object recognition



(Serre, Wolf, Bileschi, Riesenhuber and Poggio, in sub)

Examples



- The model can handle the recognition of many different object categories in complex natural images
- The model performs surprisingly well at the level of some of the best computer vision systems
- How does it compare to humans?

Roadmap

I. The model

II. Comparison with other computer vision systems

III. Comparison with human observers

Animal vs. Non-animal categ.

- Animals are rich class of stimuli
- Variety of shapes, textures
- Different depths of view, poses and sizes
- Associated with context (natural landscape)

The Stimuli

- 1,200 stimuli (from Corel database)
- 600 animals in 4 categories:
  - Head
  - Close-body
  - Medium-body
  - Far-body and groups
- 600 matched distractors (½ art., ½ nat.) to prevent reliance on low-level cues

"Head"

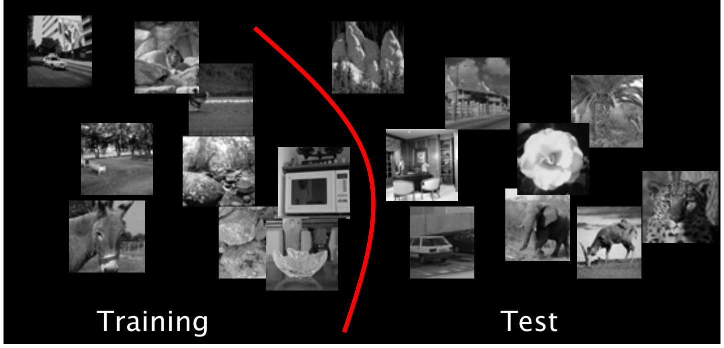
"Close-body"

"Medium-body"

"Far-body"

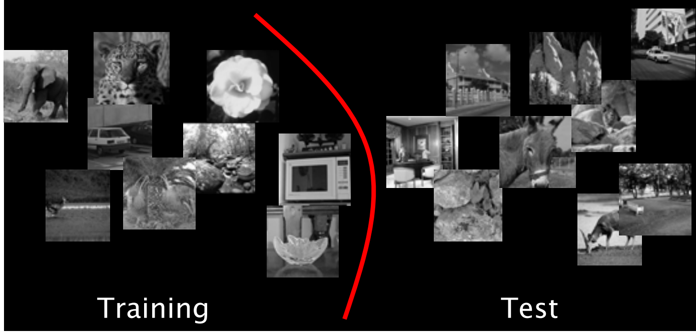
Training and testing the model

- Random splits (good estimate of expected error)
- Split 1,200 stimuli into two sets



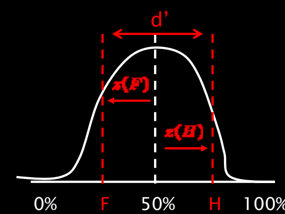
Training the model

- Repeat 20 times
- Average model performance over all

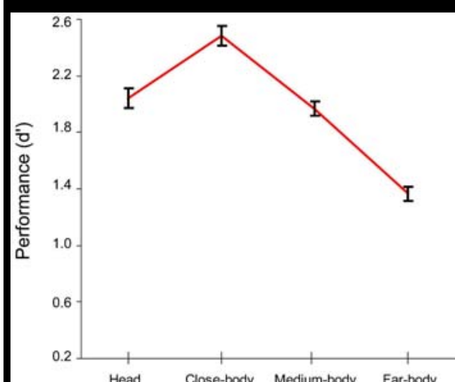


d' analysis

- Signal detection theory
  - F: false-alarm rate (non-animal images incorrectly classified as animals)
  - H: hit rate (animal images correctly classified)
  - Z: Z-score, i.e. under Gaussian assumption, how far is the rate (F or H) from chance (50%)?

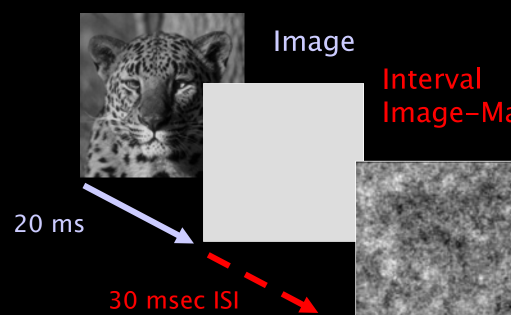


Results: Model



model

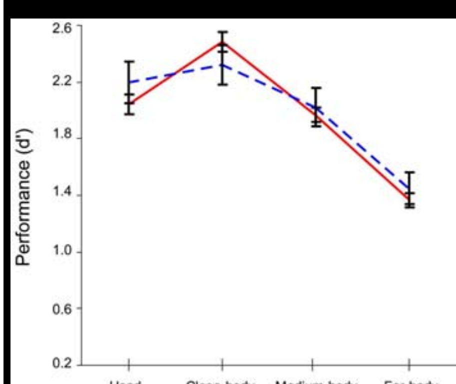
Animal vs. non-animal categ.



Animal present or not ?

(Thorpe et al, 1996; Van Rullen & Koch, 2003; Bacon-Mace et al, 2005; Oliva & Torralba, in press)

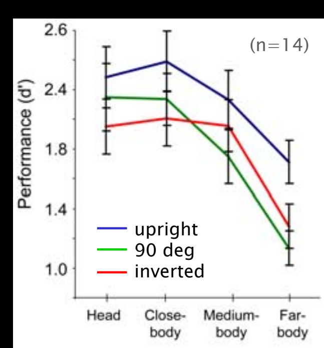
Results: Human-observers



50 ms SOA (ISI=30 ms)
model

Results: Image orientation

Human observers



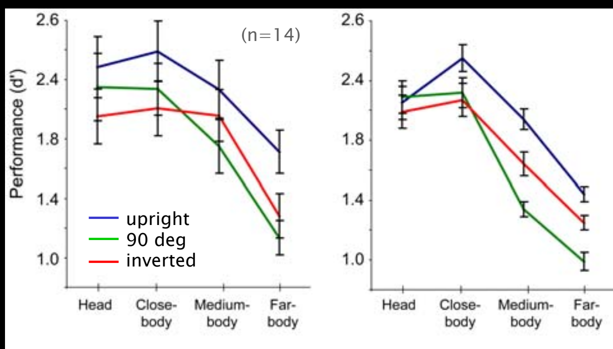
50 ms SOA (ISI=30 ms)

Robustness to image orientation is in agreement with previous results
(Rousselet et al, 2003; Guyonneau et al, ECVP 2005)

(Serre, Oliva and Poggio, in prep)

Results: Image orientation

Human observers   Model



50 ms SOA (ISI=30 ms)

(Serre, Oliva and Poggio, in prep)

Detailed comparison

- For each individual image
- How many times image classified as animal:
  - For humans: across subjects
  - For model: across 20 runs

model   humans

- Heads: ρ=0.71
- Close-body: ρ=0.84
- Medium-body: ρ=0.71
- Far-body: ρ=0.60

- The model predicts human performance extremely well when the delay between the stimulus and the mask is ~50 ms
- Under the assumption that the model correctly accounts for feedforward processing, the discrepancy for longer SOAs should be due to the cortical back-projections
- A very important question concerns the precise contribution of the feedback loops (Hochstein & Ahissar, 2002)

Contributors

- Model:
  - C. Cadieu
  - U. Knoblich
  - M. Kouh
  - G. Kreiman
  - T. Poggio
  - M. Riesenhuber
- Learning:
  - M. Giese
  - R. Liu
  - C. Koch
  - J. Louie
  - T. Poggio
  - M. Riesenhuber
  - R. Sigala
  - D. Walther
- Computer vision:
  - S. Bileschi
  - S. Chikkerur
  - E. Meyers
  - T. Poggio
  - L. Wolf
- Comparison with human-observers
  - A. Oliva
  - T. Poggio