NIPS'07 tutorial

(preliminary)

Visual Recognition in Primates and Machines

Tomaso Poggio (with Thomas Serre)

McGovern Institute for Brain Research

Center for Biological and Computational Learning

Department of Brain & Cognitive Sciences

Massachusetts Institute of Technology

Cambridge, MA 02139 USA

Motivation for studying vision:
trying to understand how the brain works

- Old dream of all philosophers and more recently of AI:
  - understand how the brain works
  - make intelligent machines

This tutorial:

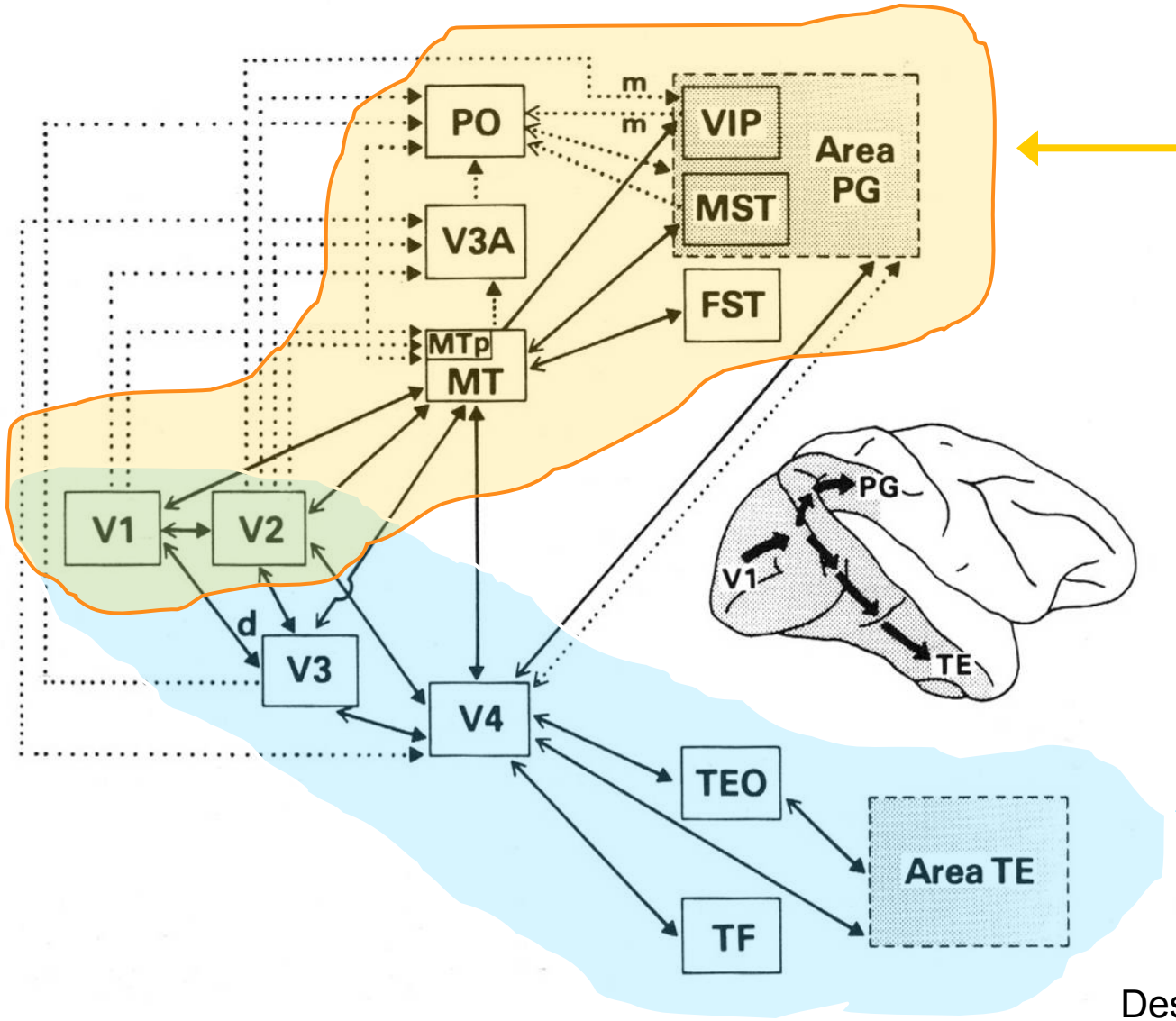
using a class of models to summarize/interpret experimental results

- Models are cartoons of reality eg Bohr's model of the hydrogen atom
- All models are "wrong"
- Some models can be useful summaries of data and some can be a good starting point for more complete theories

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

The problem: recognition in natural images (e.g., “is there an animal in the image?”)

How does visual cortex solve this problem?
How can computers solve this problem?



dorsal stream: "where"

ventral stream: "what"

Desimone & Ungerleider 1989

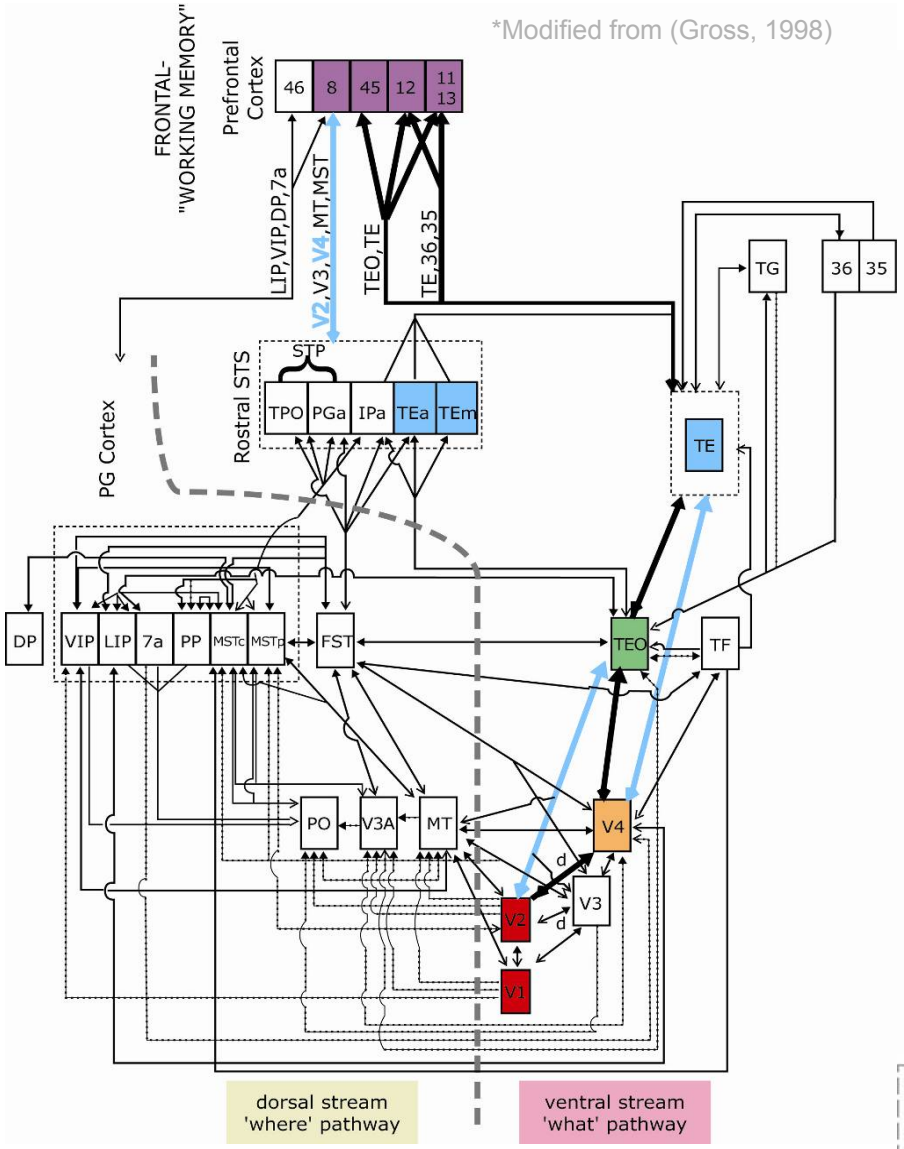
A “feedforward” version of the problem: rapid categorization

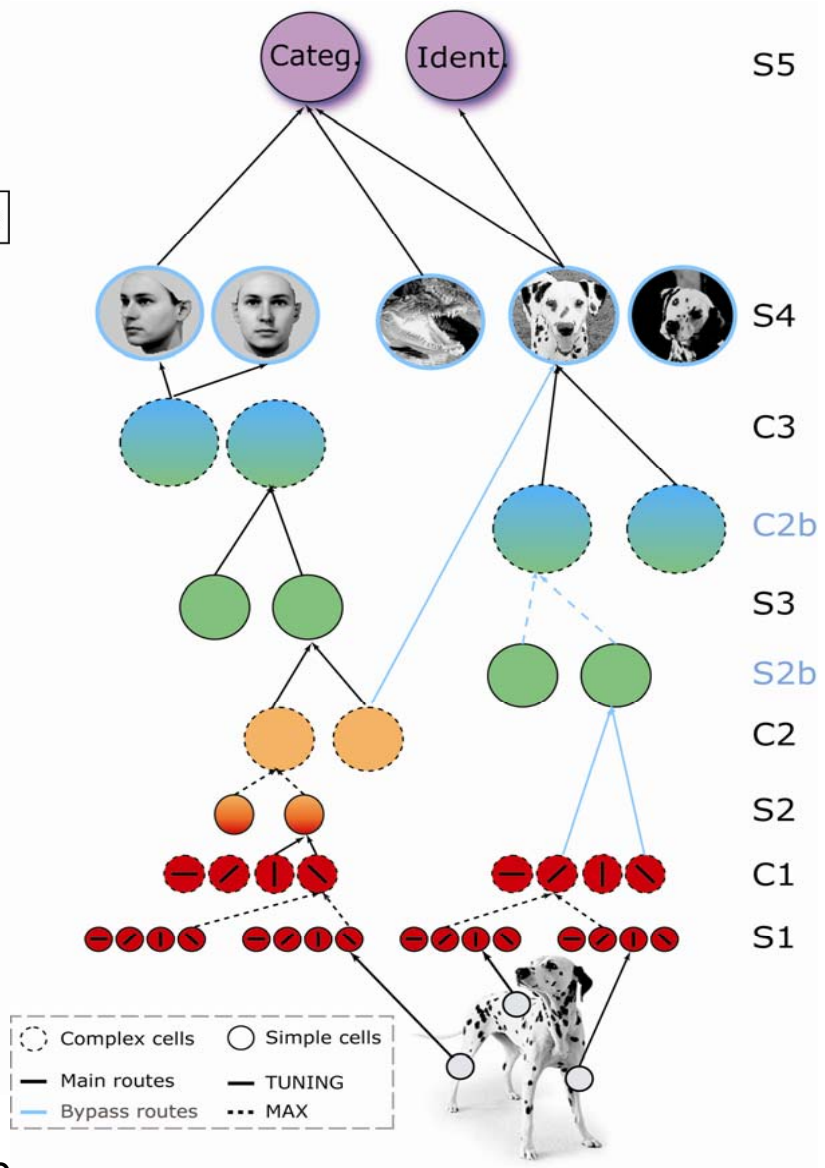
SHOW RSVP MOVIE

Movie courtesy of Jim DiCarlo

Biederman 1972; Potter 1975; Thorpe et al 1996

A model of the ventral stream which is also an algorithm





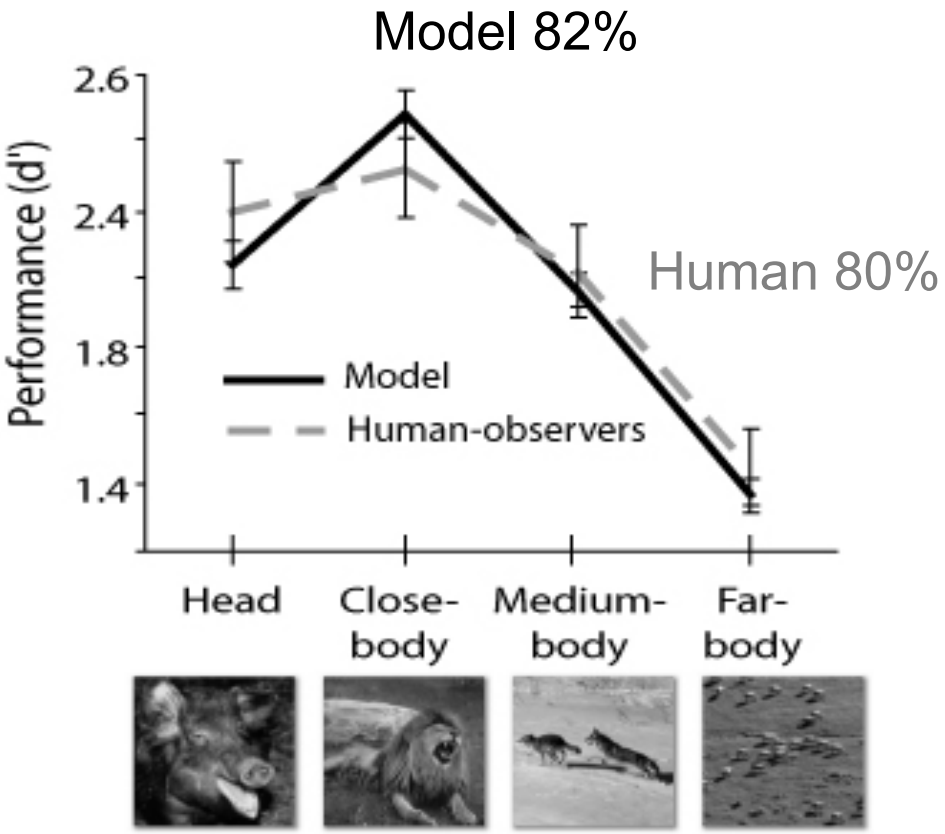
Riesenhuber & Poggio 1999, 2000;
Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005;
Serre Oliva Poggio 2007

[software available online]

...solves the problem

(if mask forces feedforward processing)...

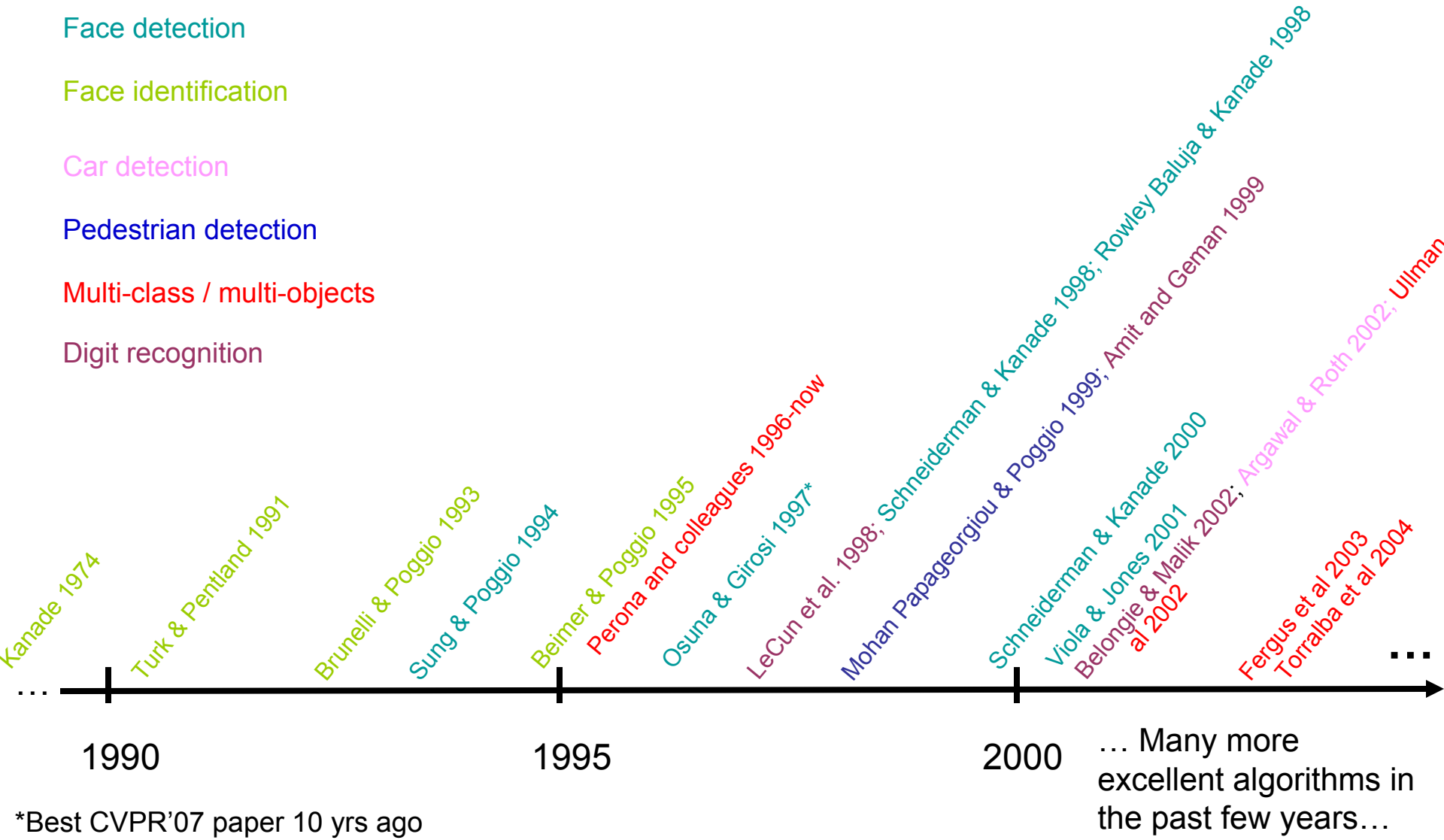
- d'~ standardized error rate
- the higher the d', the better the perf.



Serre Oliva & Poggio 2007

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

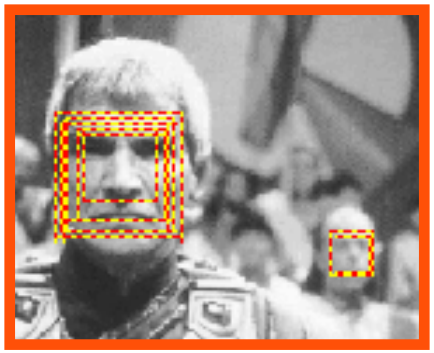
Object recognition for computer vision: personal historical perspective

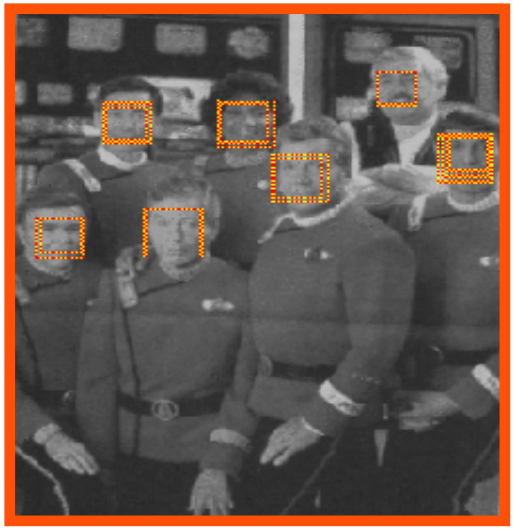


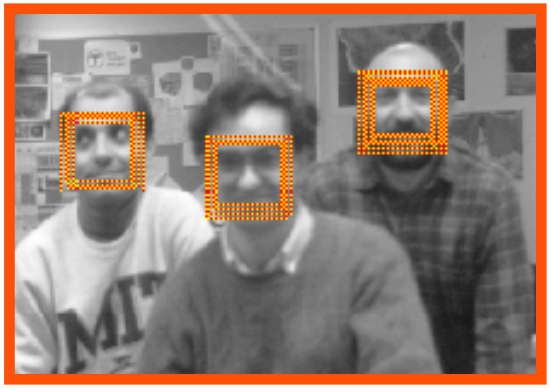
*Best CVPR'07 paper 10 yrs ago

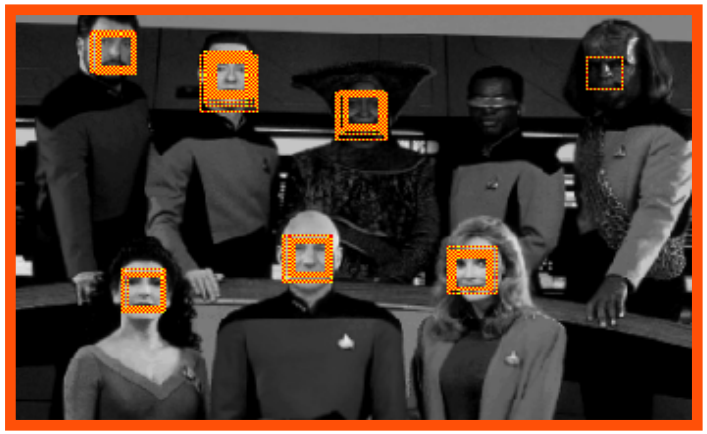
Examples: Learning Object Detection: Finding Frontal Faces

- Training Database
- 1000+ Real, 3000+ VIRTUAL
- 50,0000+ Non-Face Pattern



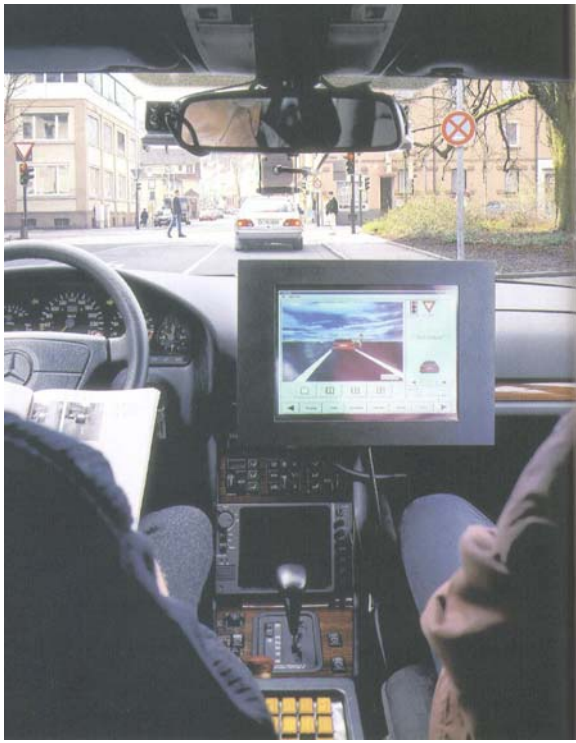






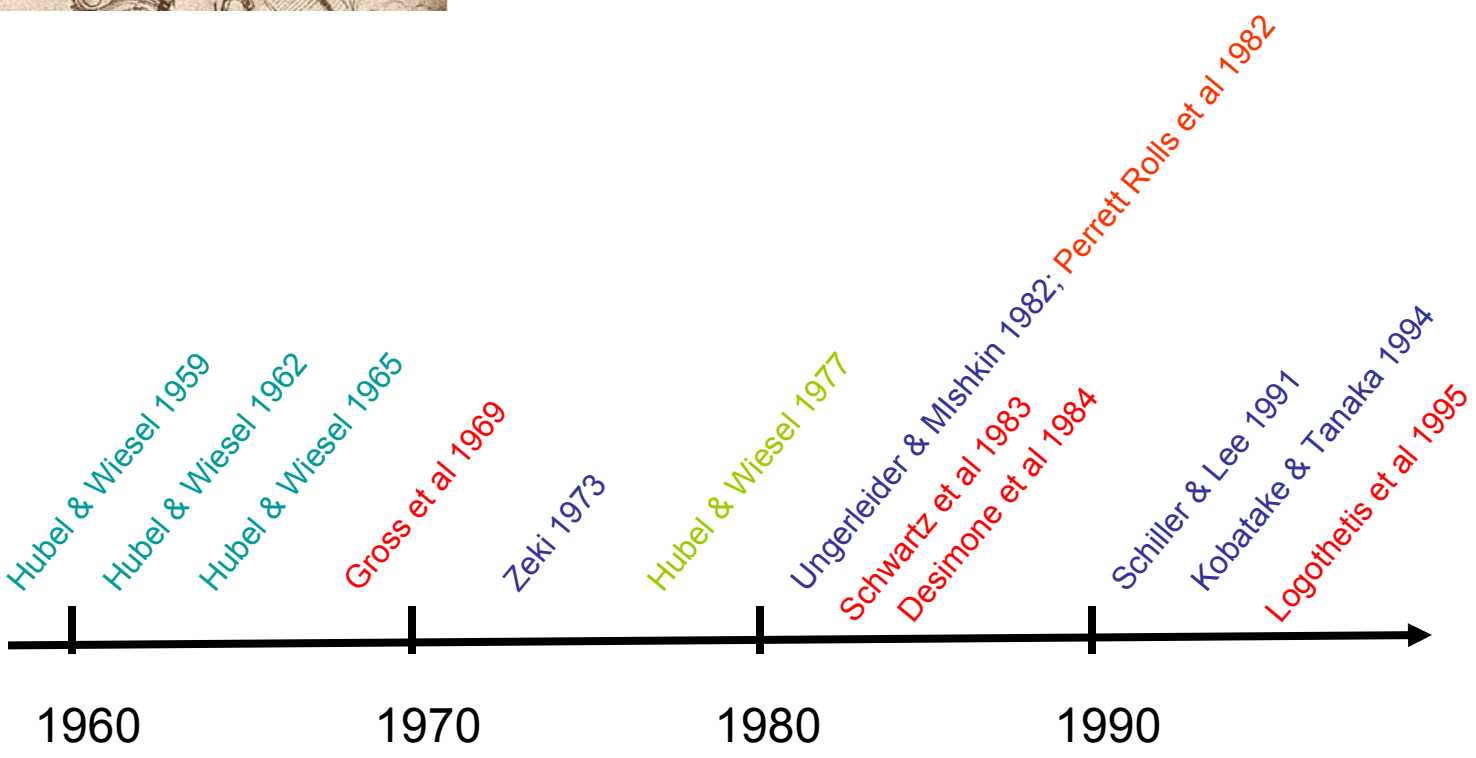
Sung & Poggio 1995

~10 year old CBCL computer vision work: pedestrian detection system in Mercedes test car now becoming a product (MobilEye)



Object recognition in cortex: Historical perspective

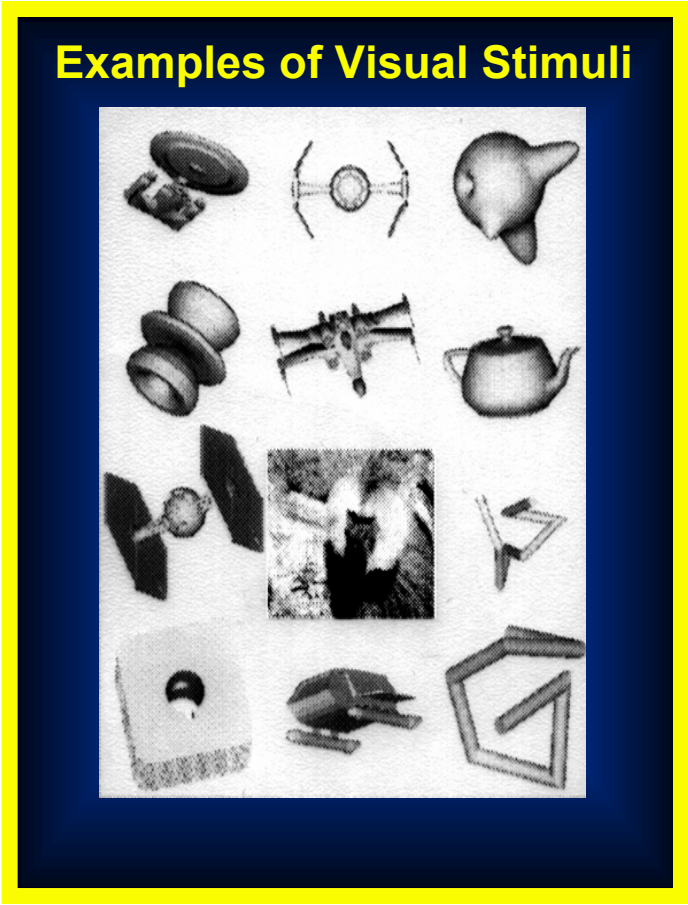
V1 cat
V1 monkey
Exstrastriate cortex
IT-STS



… Much progress in the past 10 yrs

Some personal history:

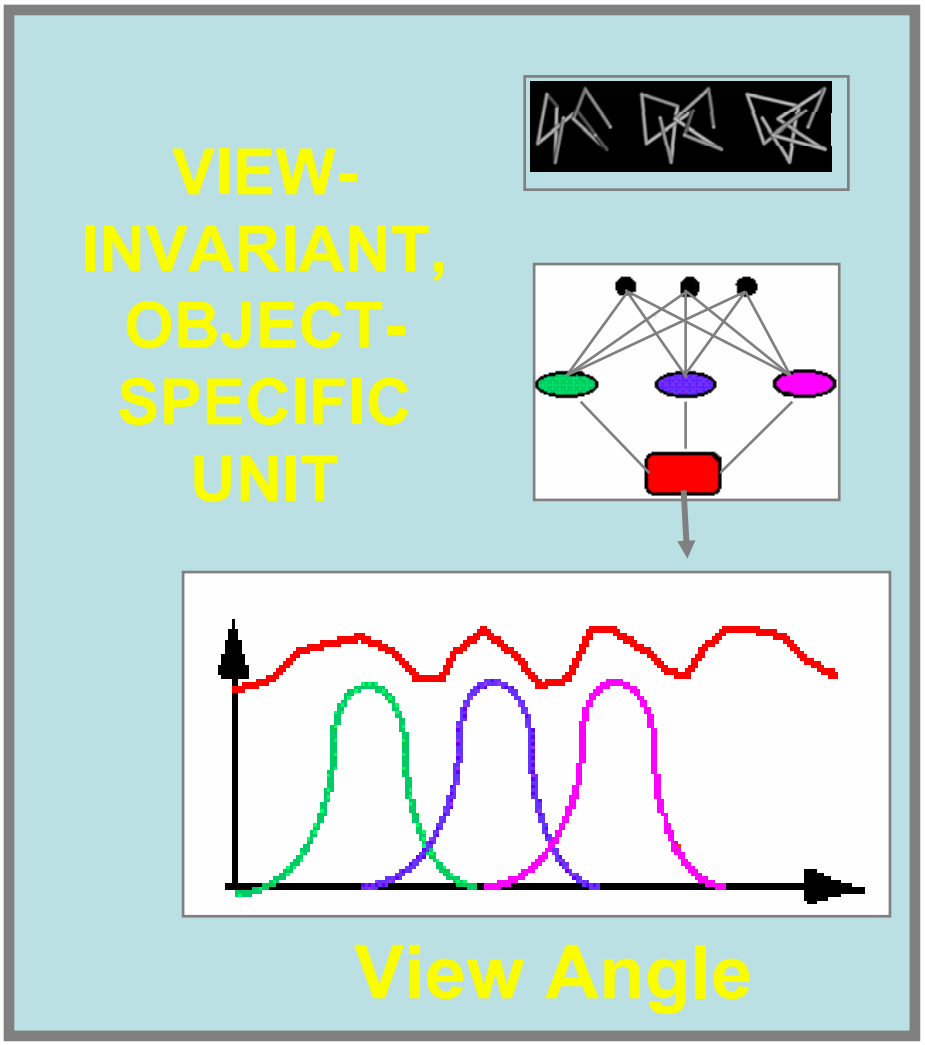
First step in developing a model: learning to recognize 3D objects in IT cortex



Poggio & Edelman 1990

An idea for a module for view-invariant identification

Architecture that accounts for invariances to 3D effects (>1 view needed to learn!)



Prediction: neurons become view-tuned through learning

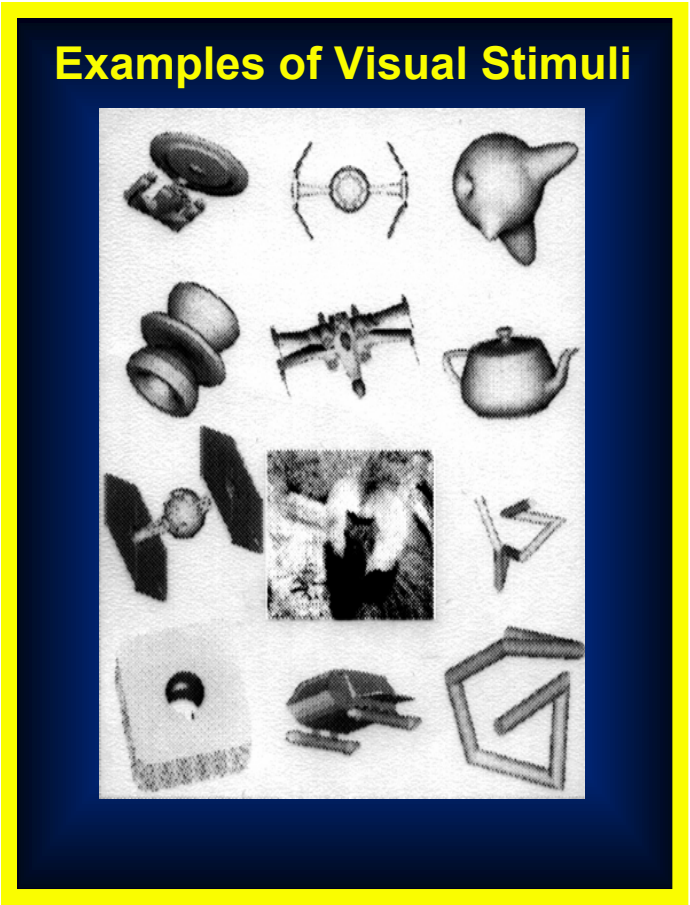
Regularization Network (GRBF) with Gaussian kernels

Poggio & Edelman 1990

Learning to Recognize 3D Objects in IT Cortex

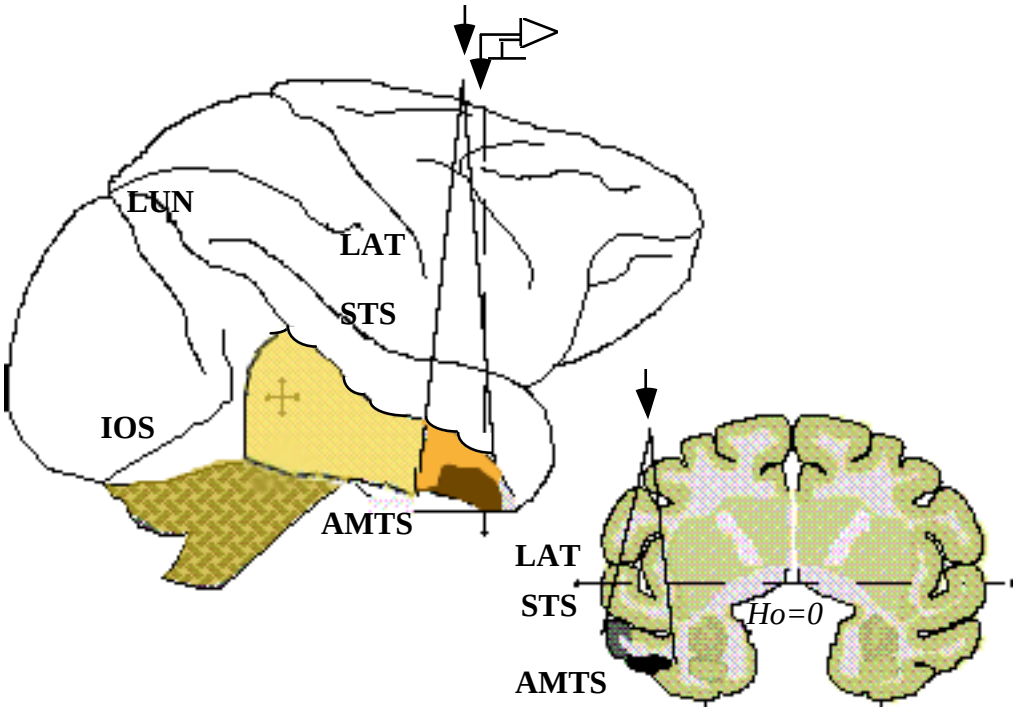
After human psychophysics (Buelthoff, Edelman, Tarr, Sinha, …), which supports models based on view-tuned units...

… physiology!



Logothetis Pauls & Poggio 1995

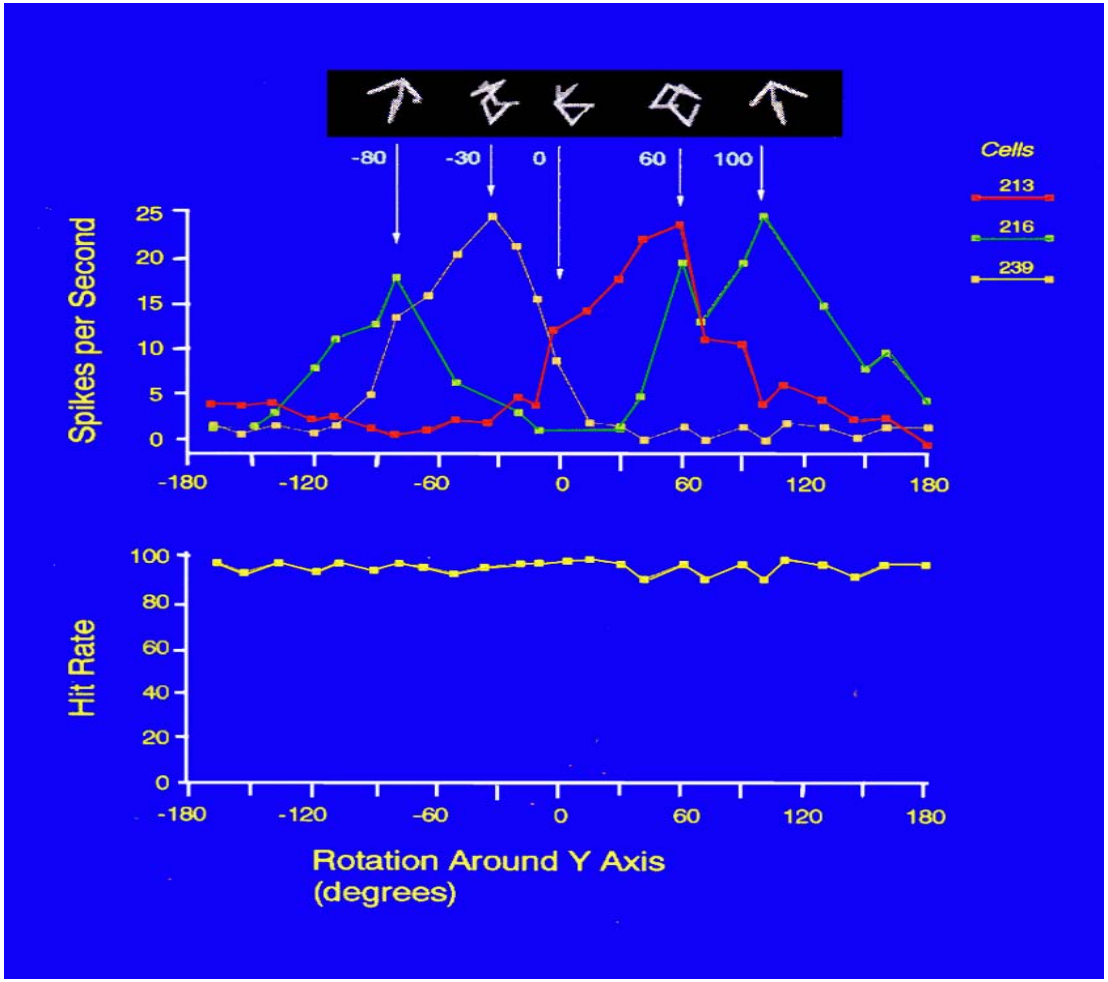
Recording Sites in Anterior IT



…neurons tuned to faces are intermingled nearby….

Logothetis, Pauls & Poggio 1995

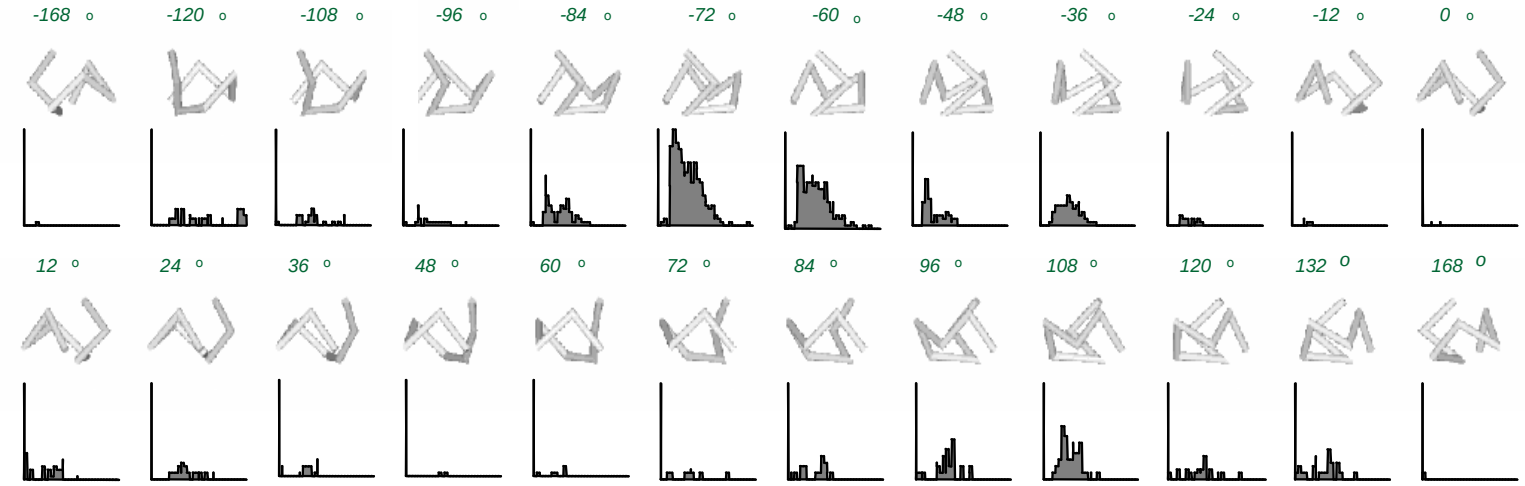
Neurons tuned to object views as predicted by model



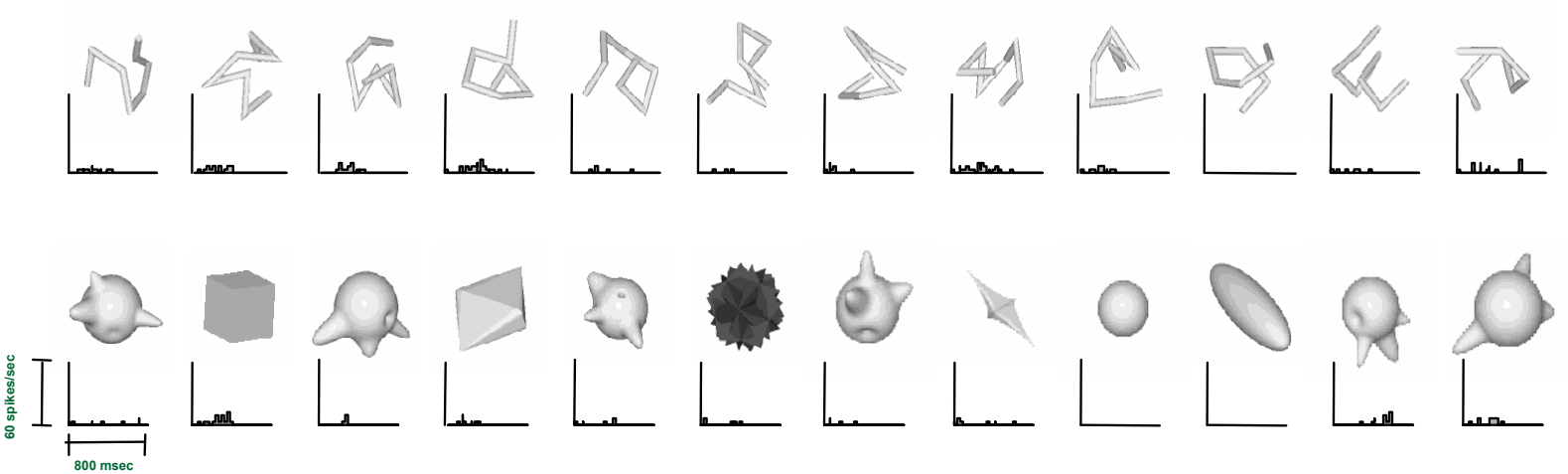
Logothetis Pauls & Poggio 1995

A “View-Tuned” IT Cell

Target Views

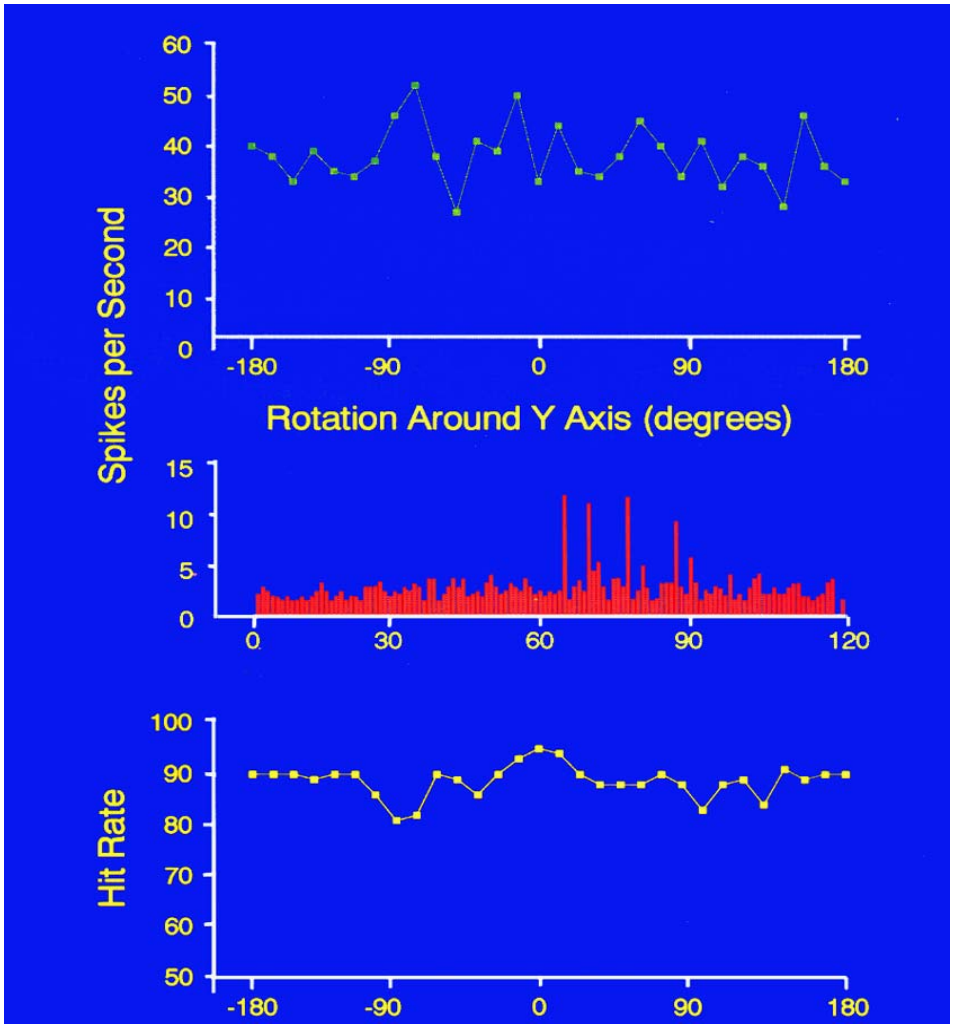


Distractors



Logothetis Pauls & Poggio 1995

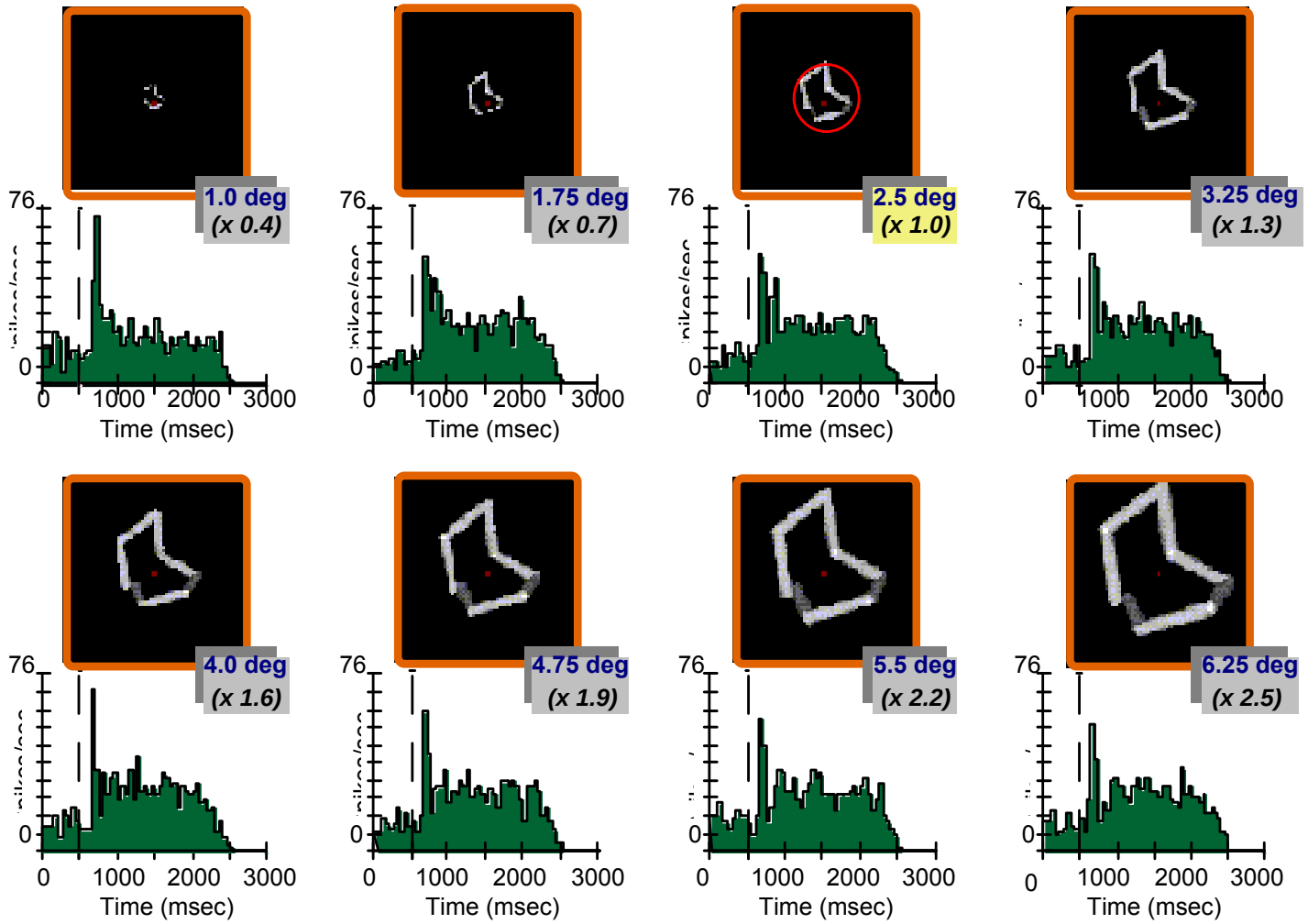
But also view-invariant object-specific neurons (5 of them over 1000 recordings)



Logothetis Pauls & Poggio 1995

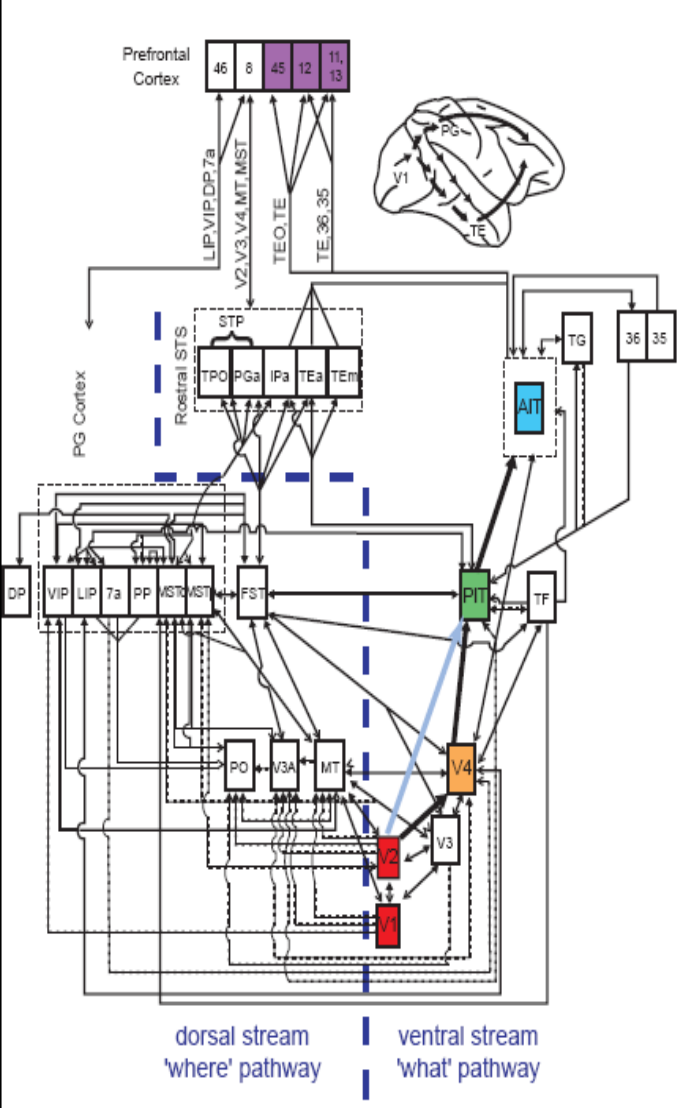
View-tuned cells:

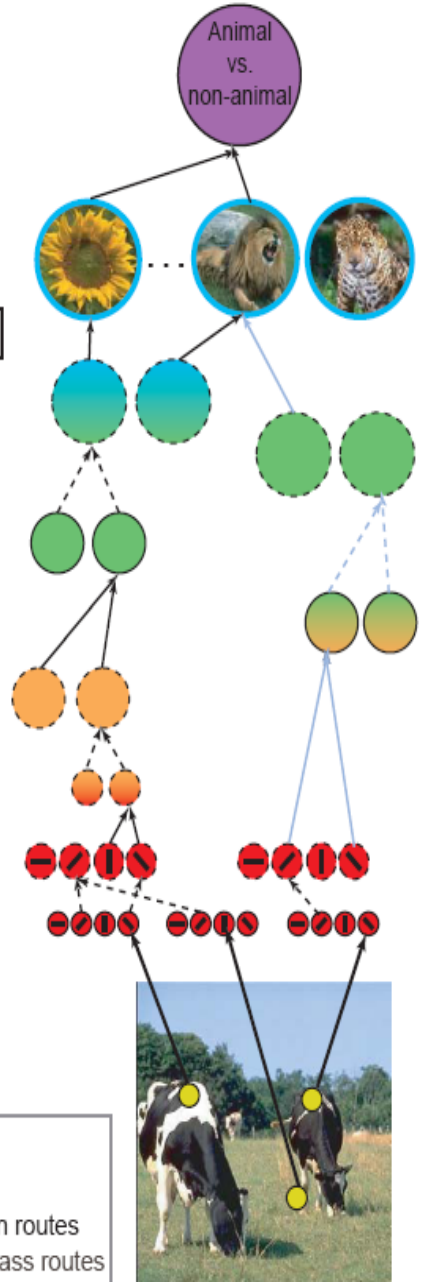
scale invariance (one training view only) motivates present model

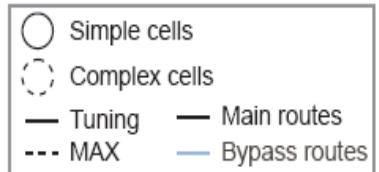


Logothetis Pauls & Poggio 1995

From “HMAX” to the model now…







Model layers	Corresponding brain area (tentative)	RF sizes	Number units	
classifier	PFC		$1.0\ 10^0$	
S4	AIT	>4.4°	$1.5\ 10^2$	~ 5,000 subunits
C3	PIT - AIT	>4.4°	$2.5\ 10^3$	
C2b	PIT	>4.4°	$2.5\ 10^3$	
S3	PIT	1.2°- 3.2°	$7.4\ 10^4$	~ 100 subunits
S2b	V4 - PIT	0.9°- 4.4°	$1.0\ 10^7$	~ 100 subunits
C2	V4	1.1°- 3.0°	$2.8\ 10^5$	
S2	V2 - V4	0.6°- 2.4°	$1.0\ 10^7$	~ 10 subunits
C1	V1 - V2	0.4°- 1.6°	$1.2\ 10^4$	
S1	V1 - V2	0.2°- 1.1°	$1.6\ 10^6$	

Supervised task-dependent learning

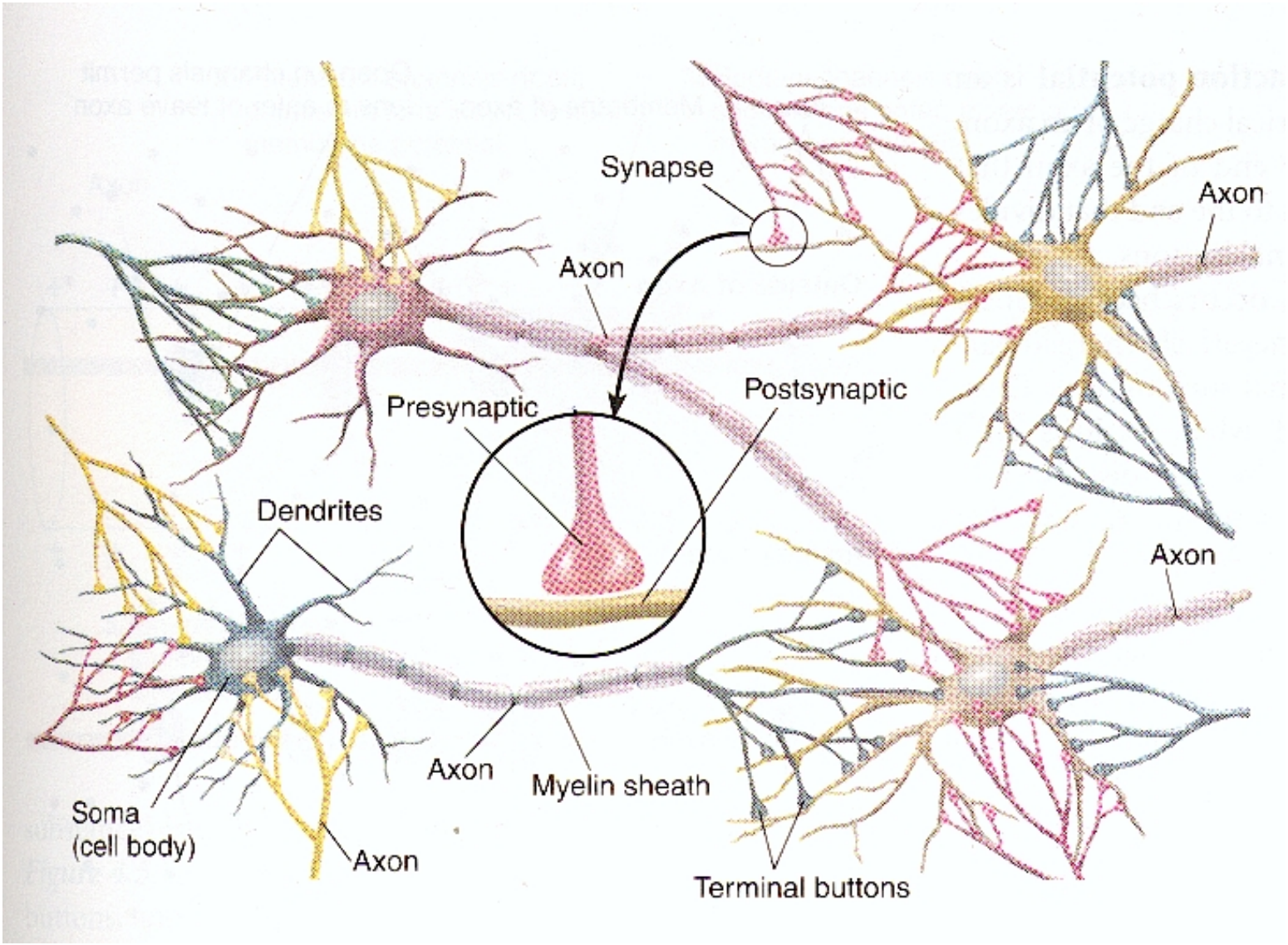
Unsupervised task-independent learning

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

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

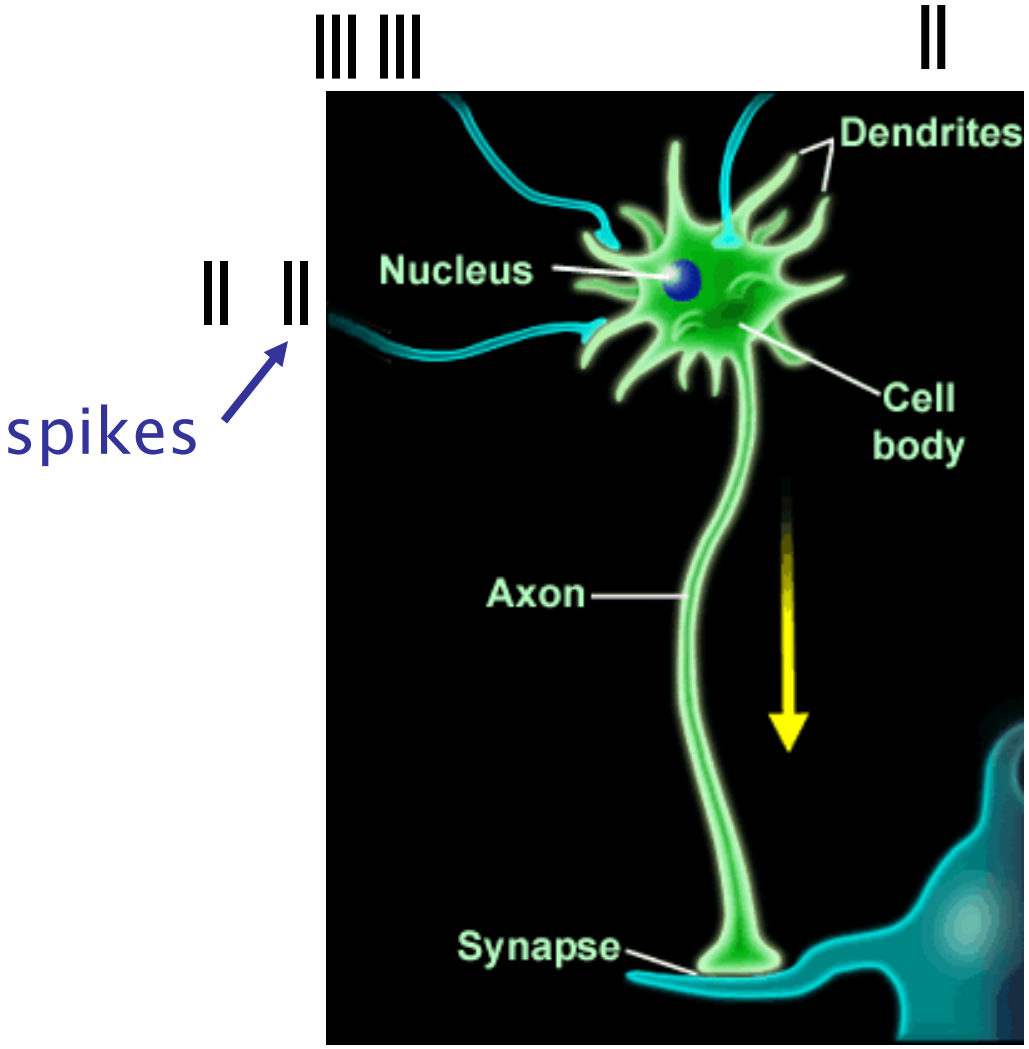
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

Neural Circuits



Source: Modified from Jody Culham's web slides

Neuron basics



INPUT= pulses or graded potentials

COMPUTATION = Analog

OUTPUT = Chemical

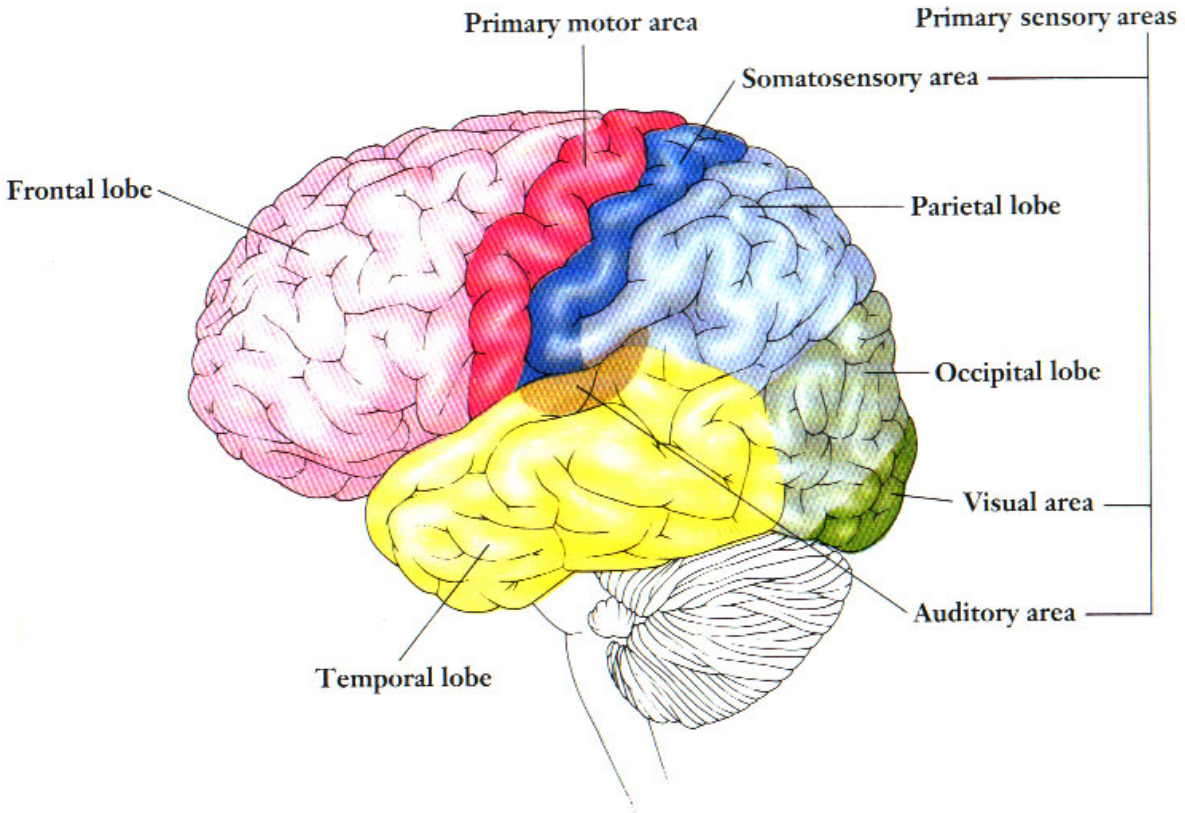
Some numbers

- Human Brain
  - 10^{11}-10^{12} neurons (1 million flies ☺)
  - 10^{14}- 10^{15} synapses
- Neuron
  - Fundamental space dimension:
    - fine dendrites : 0.1 μ diameter; lipid bilayer membrane : 5 nm thick; specific proteins : pumps, channels, receptors, enzymes
  - Fundamental time length : 1 msec

The cerebral cortex

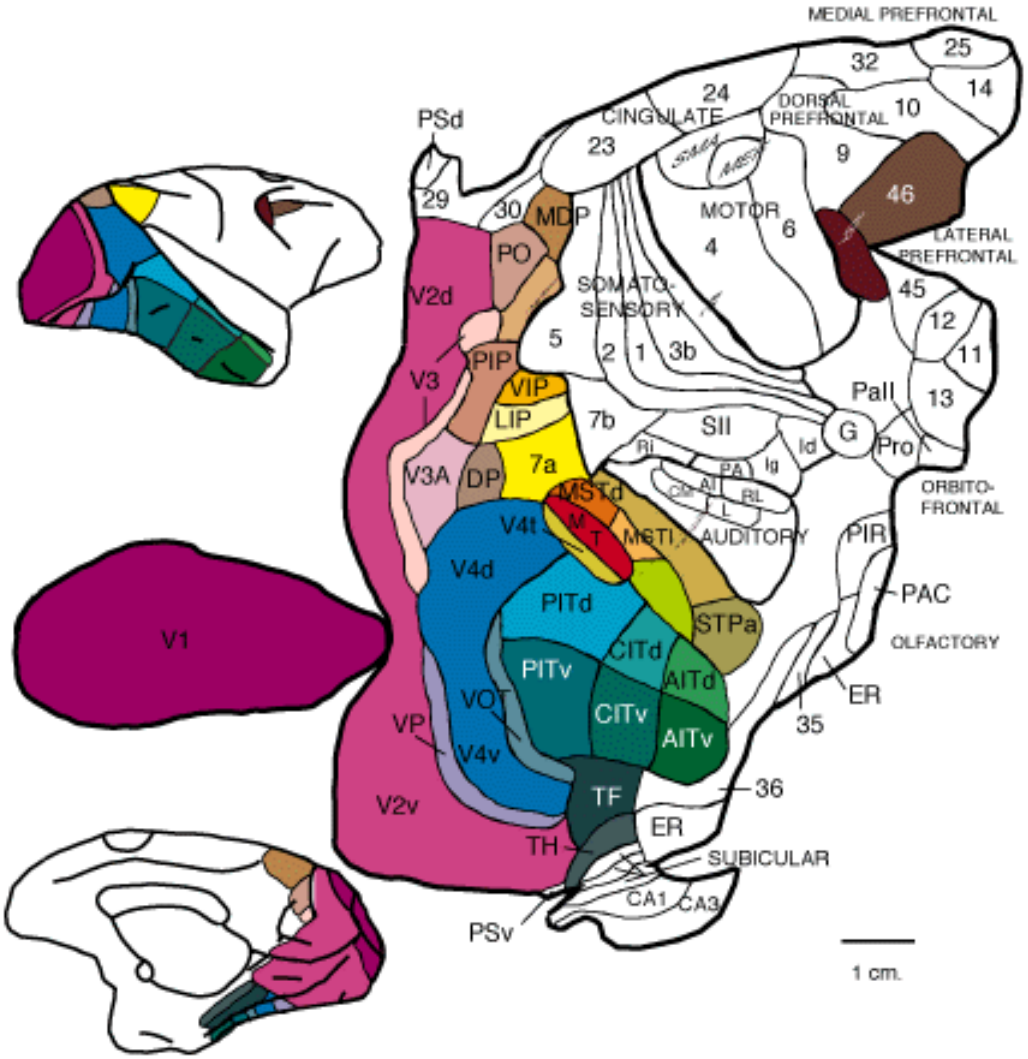
	Human	**Macaque**
Thickness	3 – 4 mm	1 – 2 mm
Total surface area (both sides)	~1600 cm2 (~50cm diam)	~160 cm2 (~15cm diam)
Neurons /mm²	~10^5/ mm2	~ 10^5/ mm2
Total cortical neurons	~2 x 1010	~2 x 109
Visual cortex	300 – 500 cm2	80+cm2
Visual Neurons	~4 x 109	~109 neurons

Gross Brain Anatomy



A large percentage of the cortex devoted to vision

The Visual System



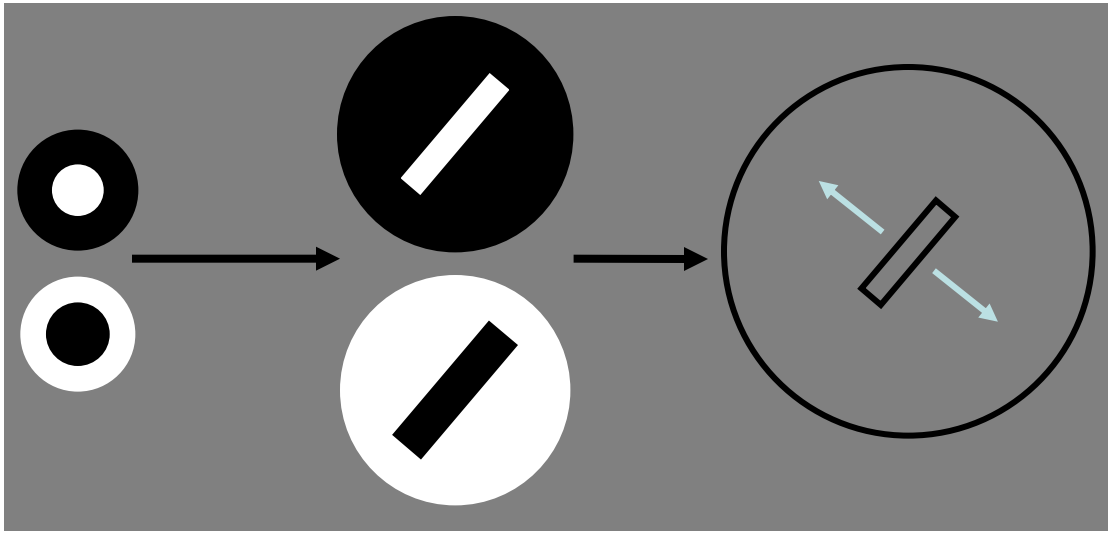
[Van Essen & Anderson, 1990]

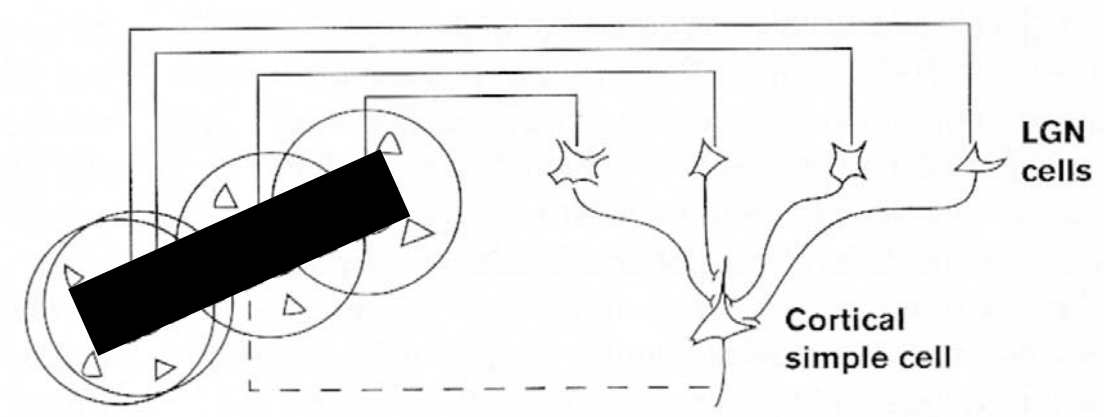
V1: hierarchy of simple and complex cells

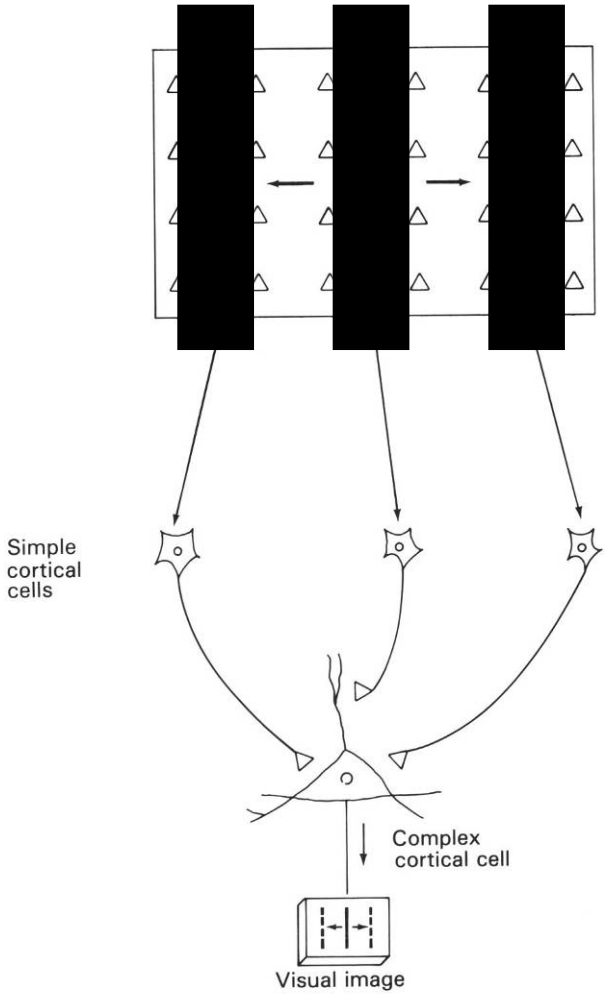
LGN-type cells

Simple cells

Complex cells





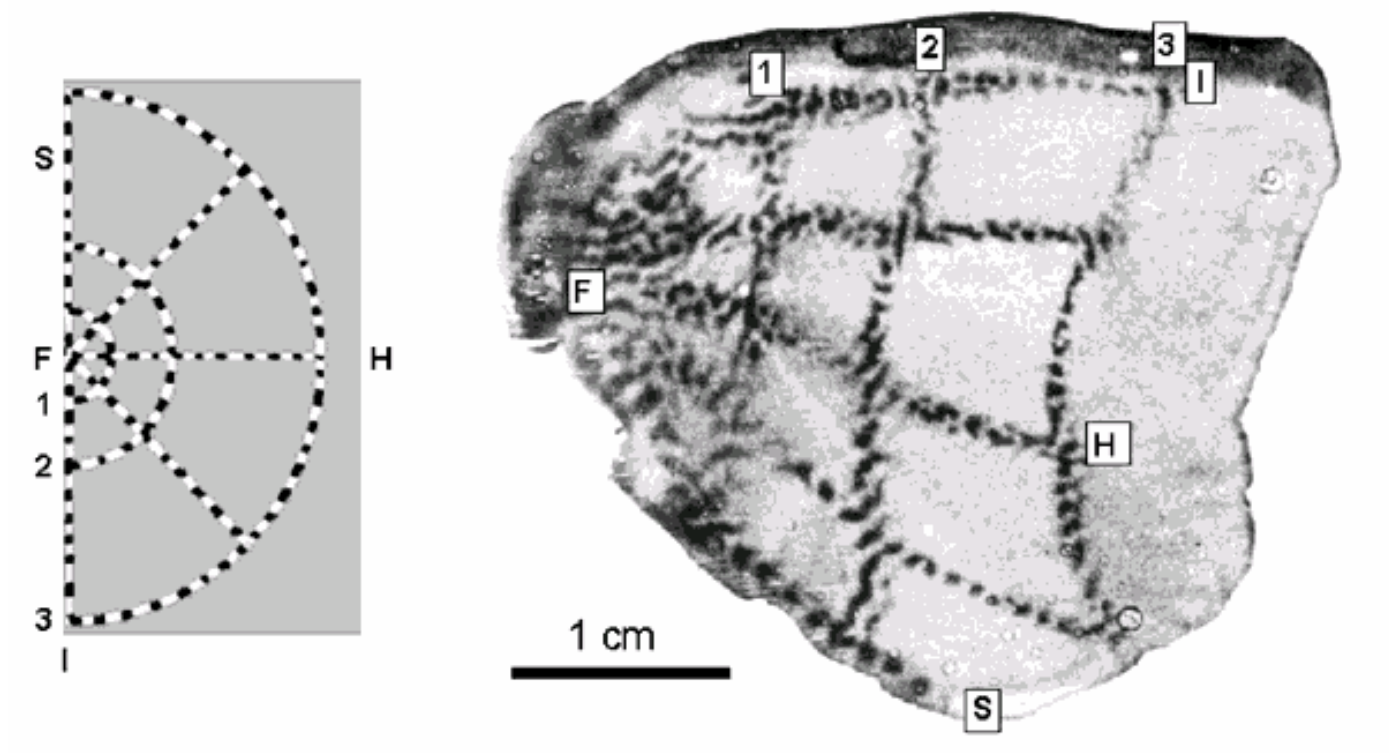


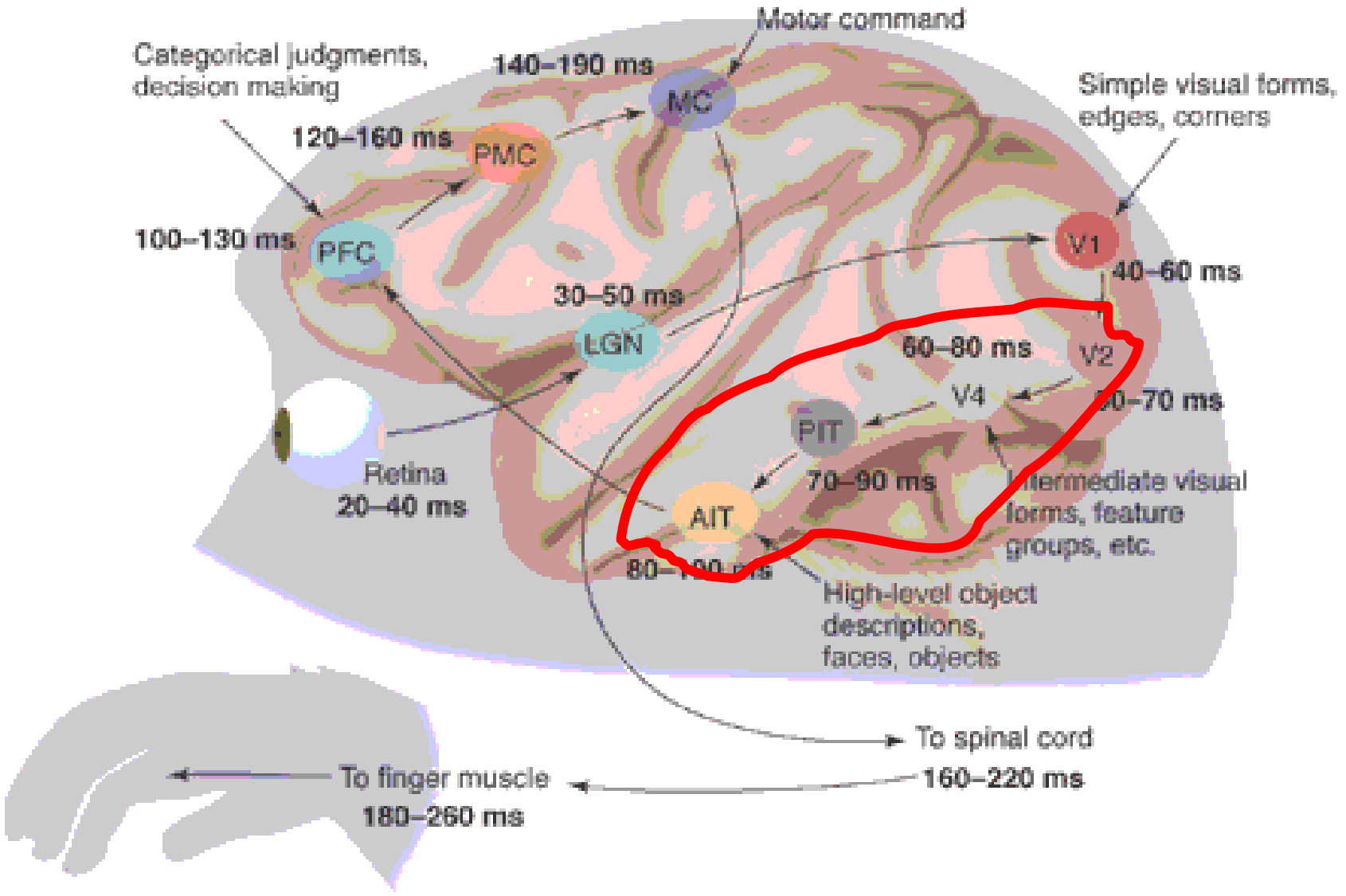
(Hubel & Wiesel 1959)

V1: Orientation selectivity

Hubel & Wiesel
movie

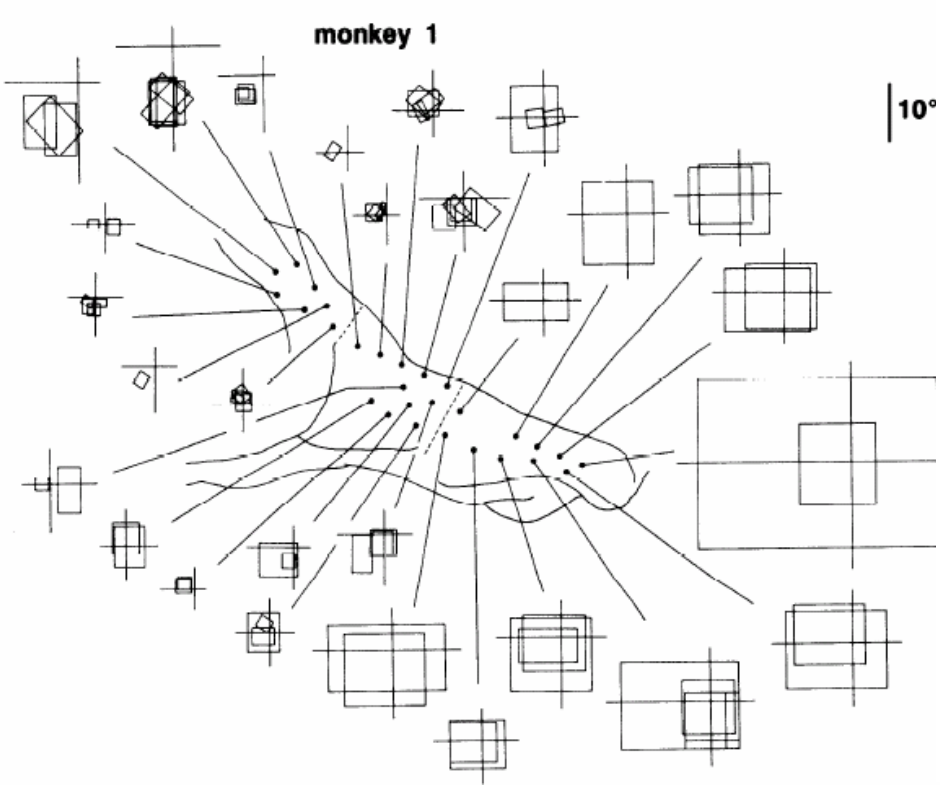
V1: Retinotopy



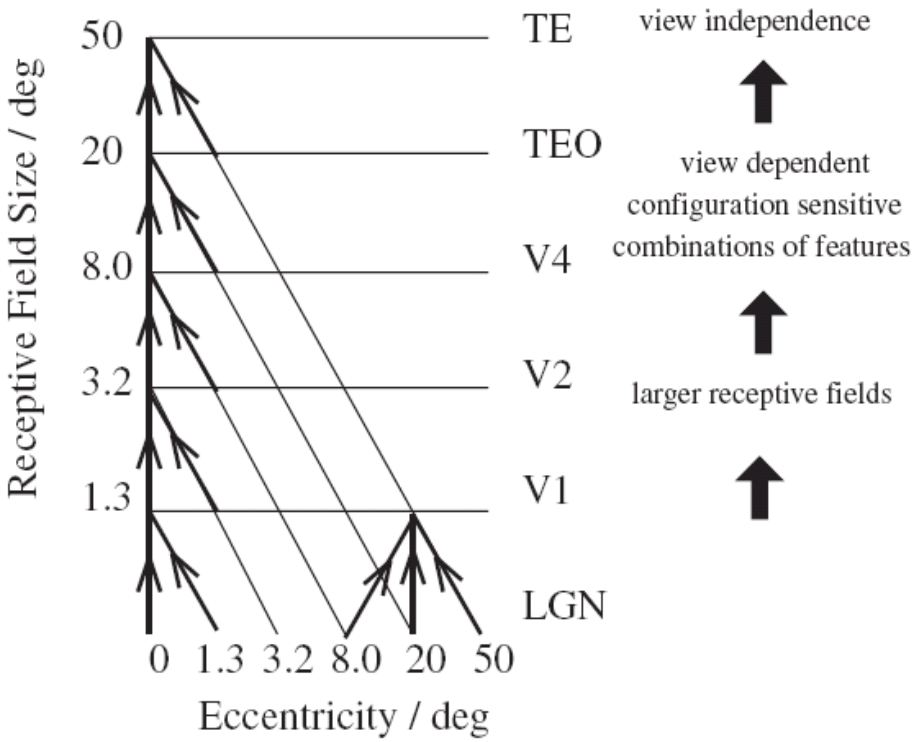


(Thorpe and Fabre-Thorpe, 2001)

Beyond V1: A gradual increase in RF size

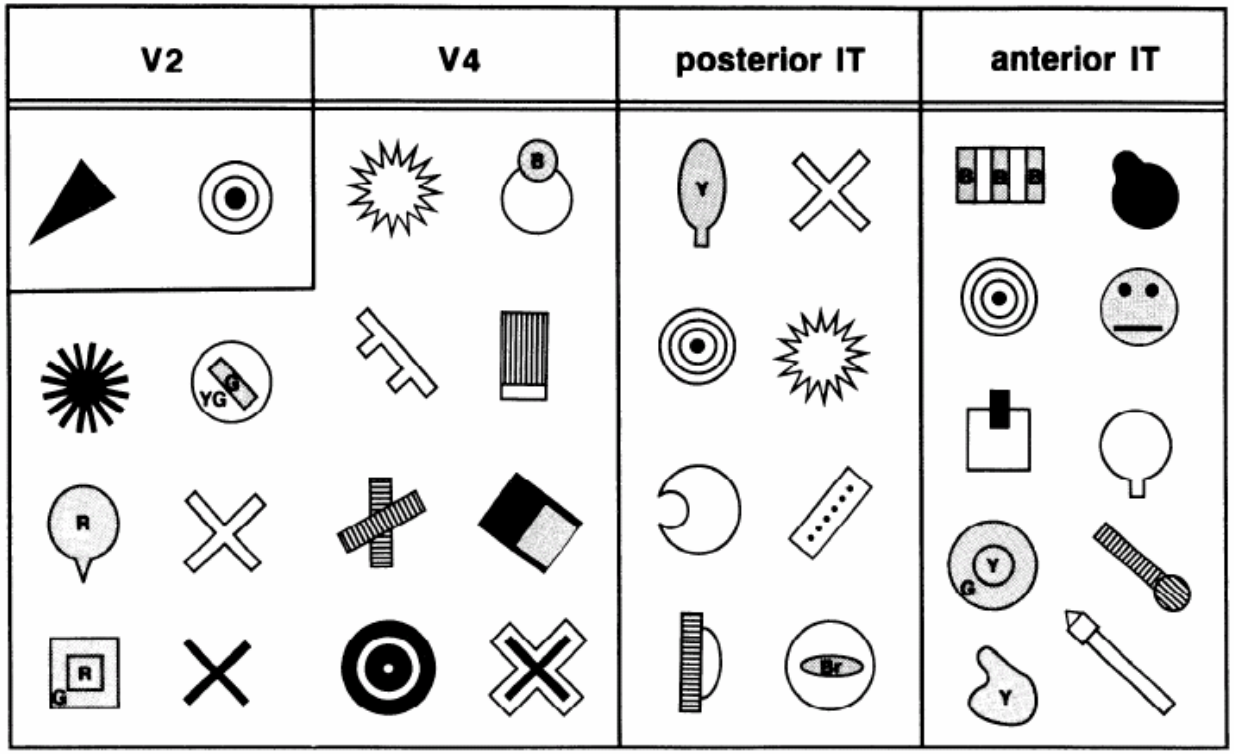


Reproduced from [Kobatake & Tanaka, 1994]



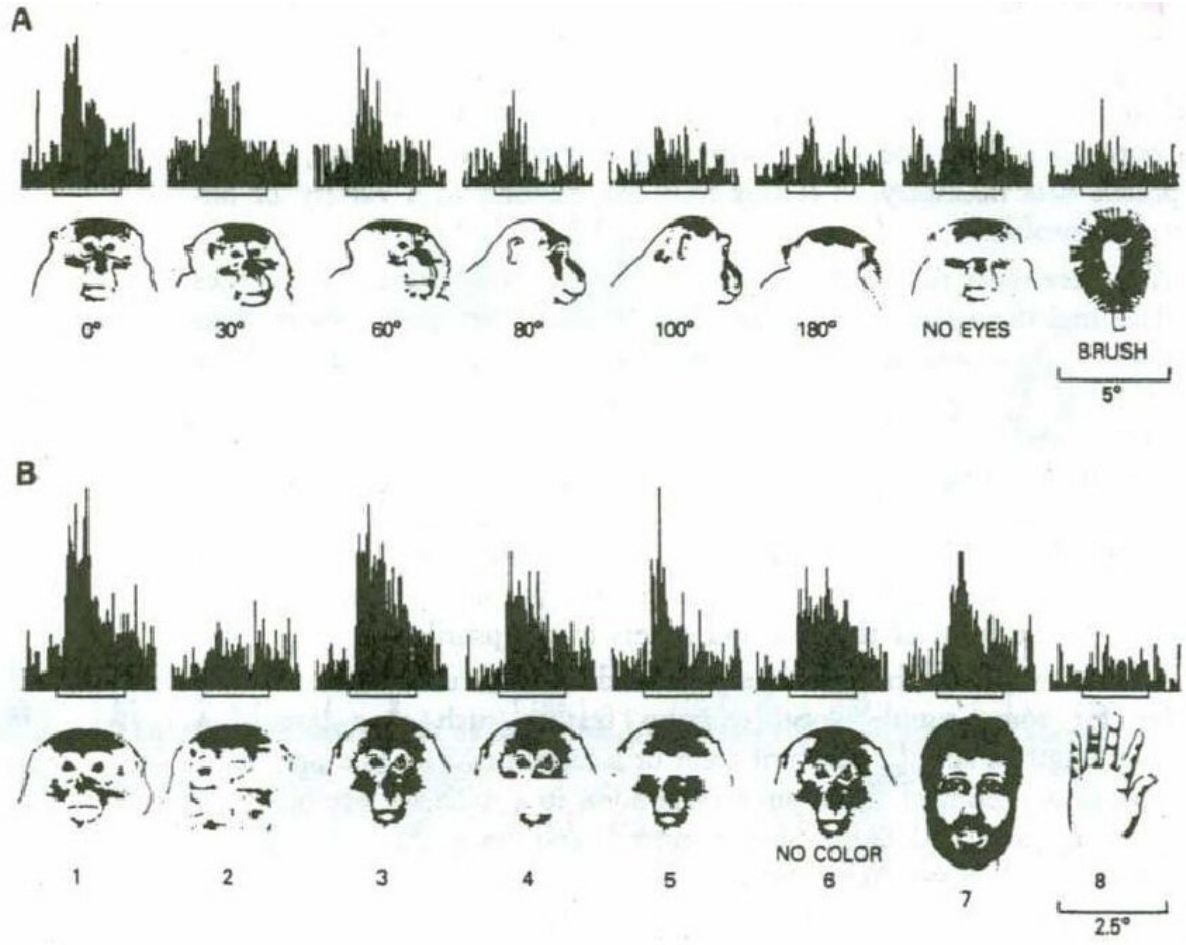
Reproduced from [Rolls, 2004]

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



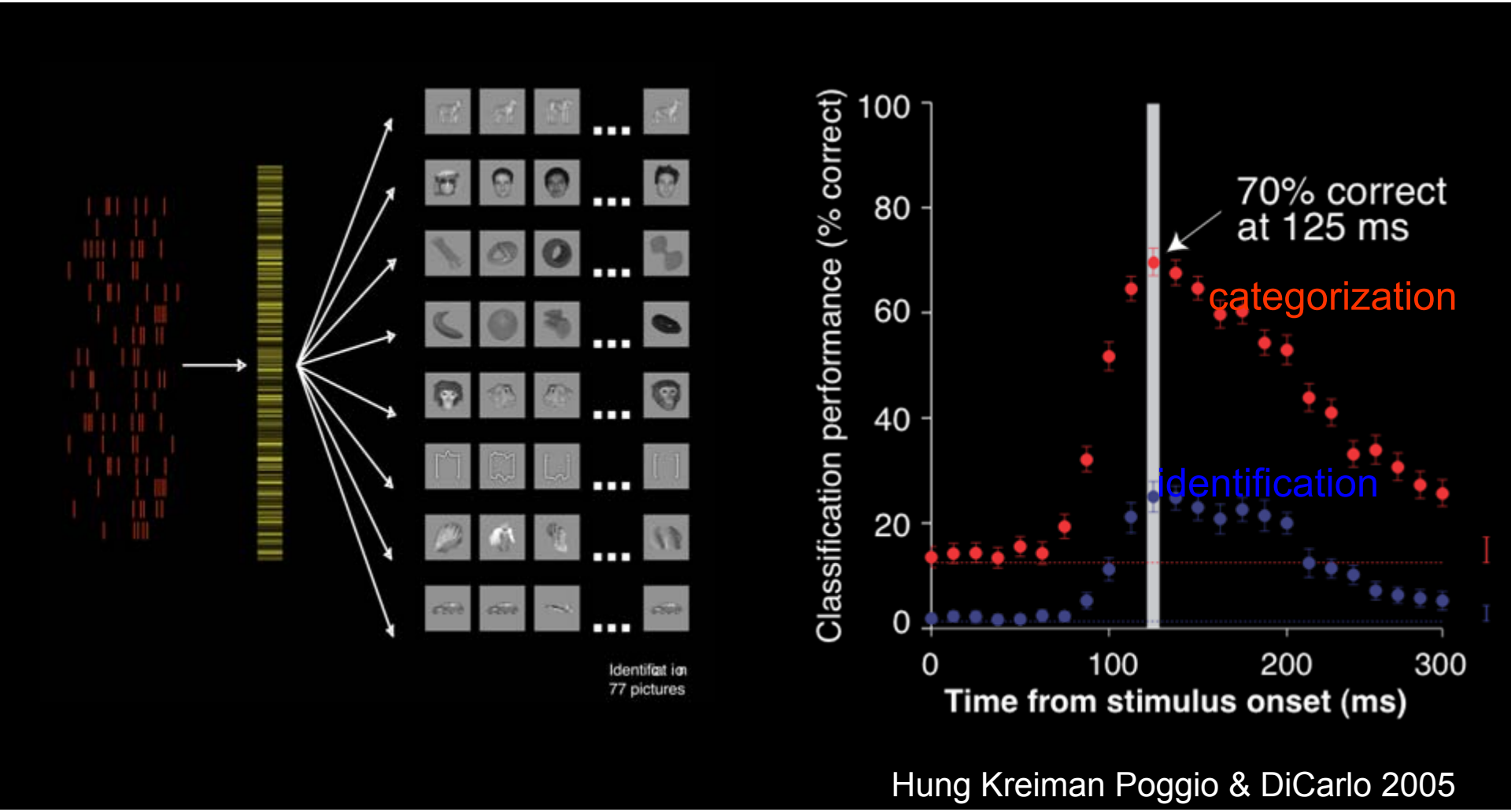
Reproduced from (Kobatake & Tanaka, 1994)

AIT: Face cells



Reproduced from (Desimone et al. 1984)

AIT: Immediate recognition

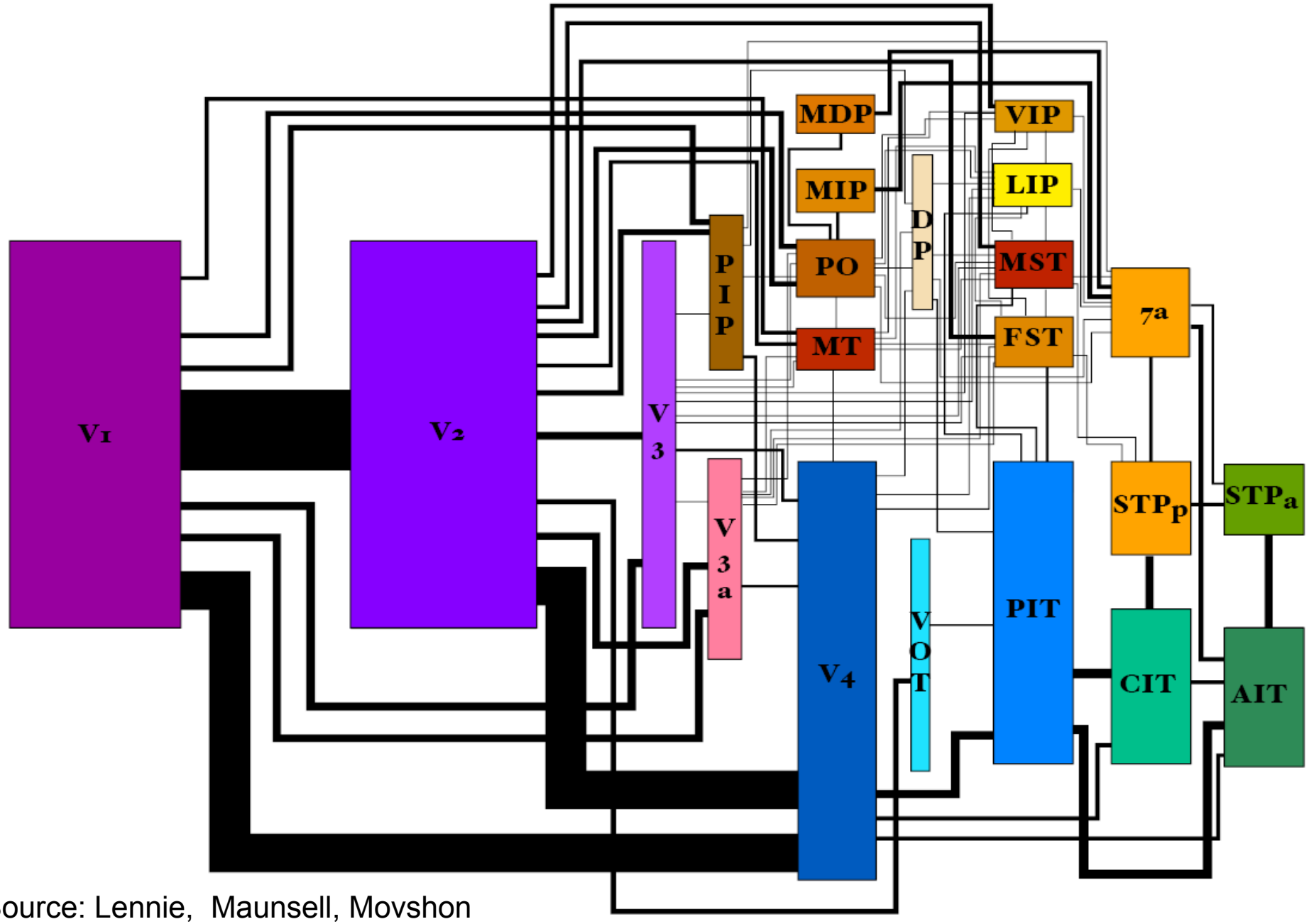


Hung Kreiman Poggio & DiCarlo 2005

See also Oram & Perrett 1992; Tovee et al 1993; Celebrini et al 1993; Ringach et al 1997; Rolls et al 1999; Keysers et al 2001

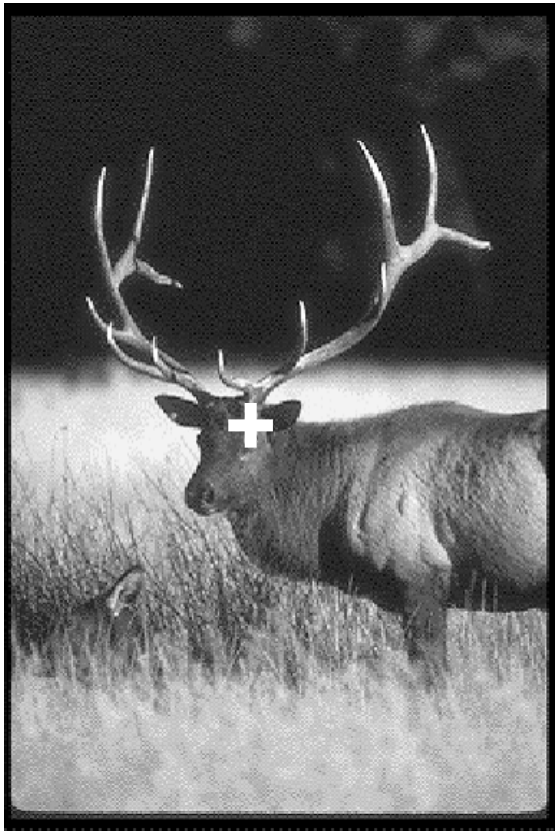
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

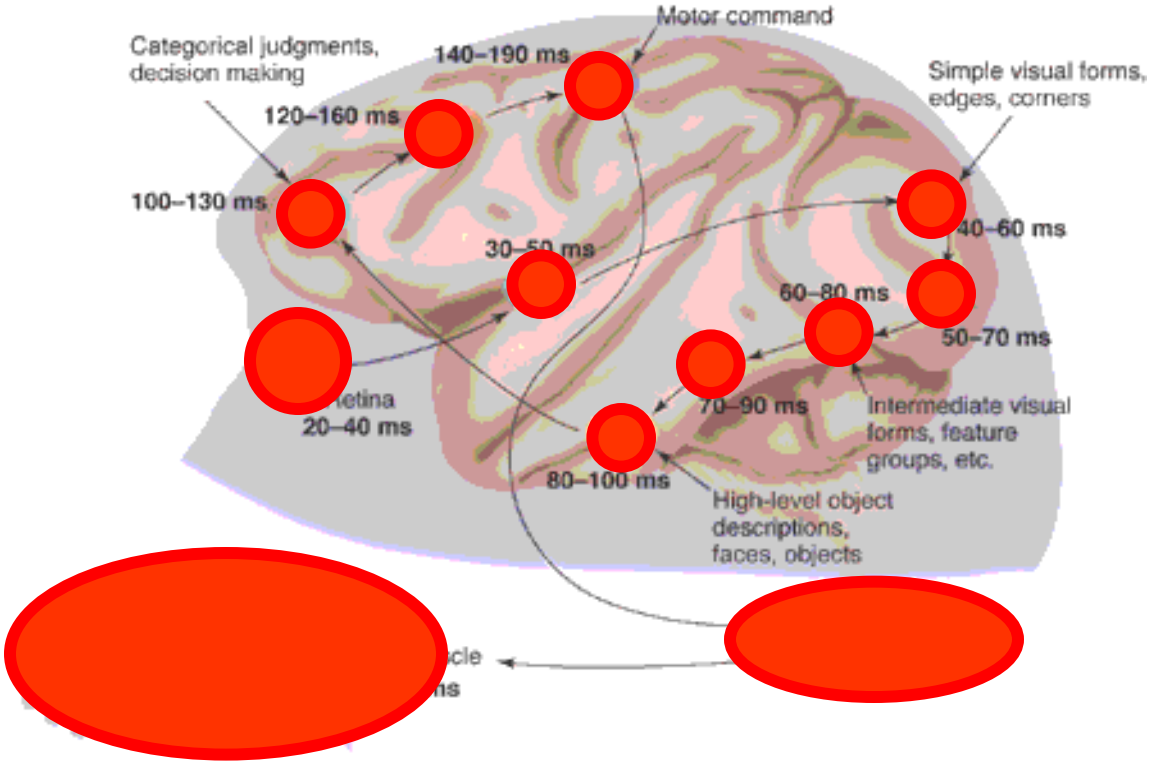
The ventral stream



Source: Lennie, Maunsell, Movshon

We consider feedforward architecture only



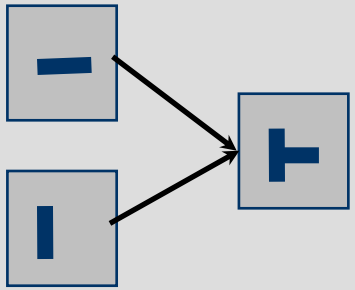
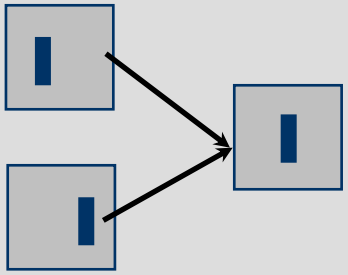


(Thorpe and Fabre-Thorpe, 2001)

Our present model of the ventral stream: feedforward, accounting only for "immediate recognition"

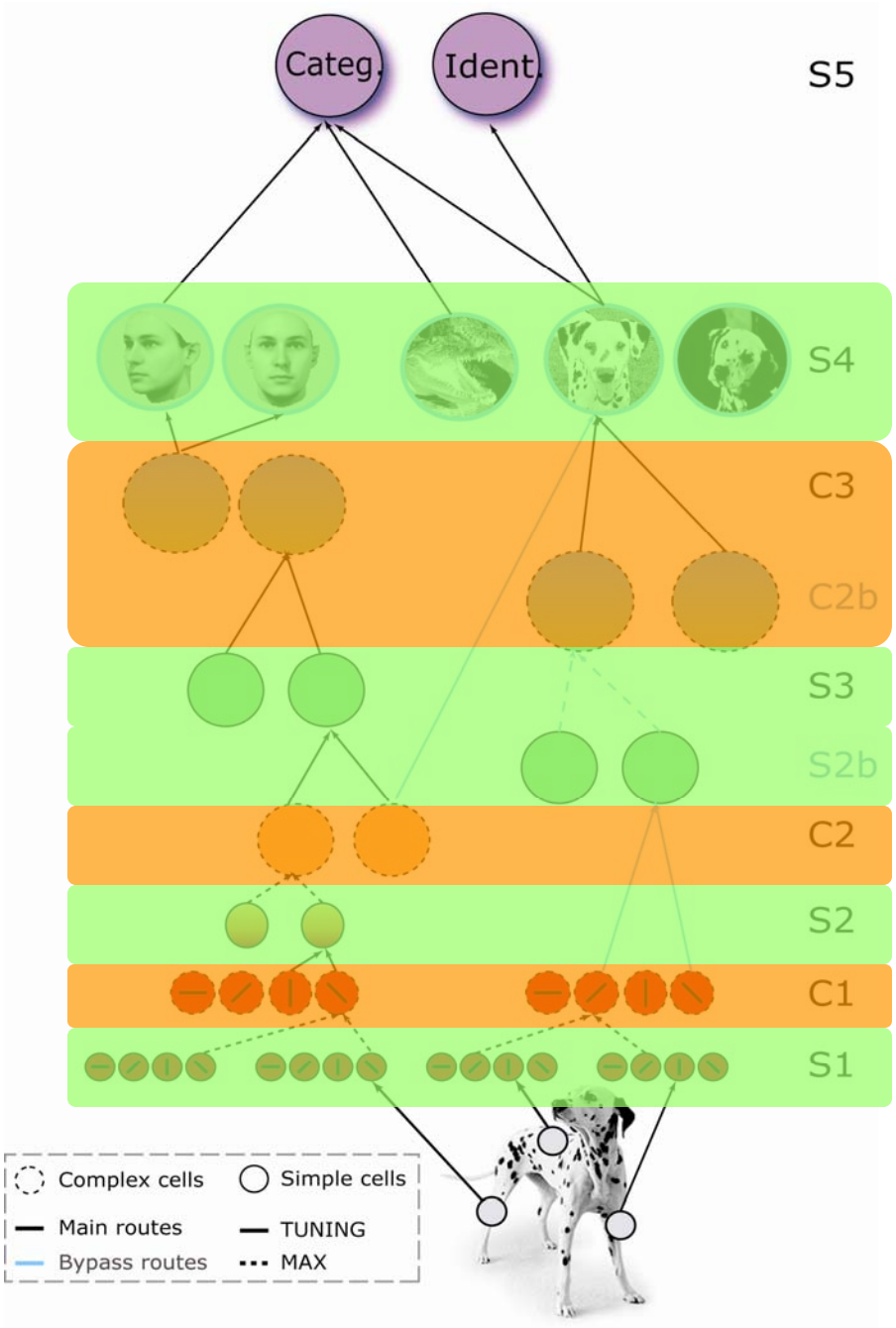
- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998; Amit & Mascaro 2003; Deco & Rolls 2006…)
- As a biological model of object recognition in the ventral stream it is *perhaps* the most quantitative and faithful to known biology (though many details/facts are unknown or still to be incorporated)

Two key computations

Unit types	Pooling	Computation	Operation
Simple		Selectivity / template matching	Gaussian-tuning / and-like
Complex		Invariance	Soft-max / or-like

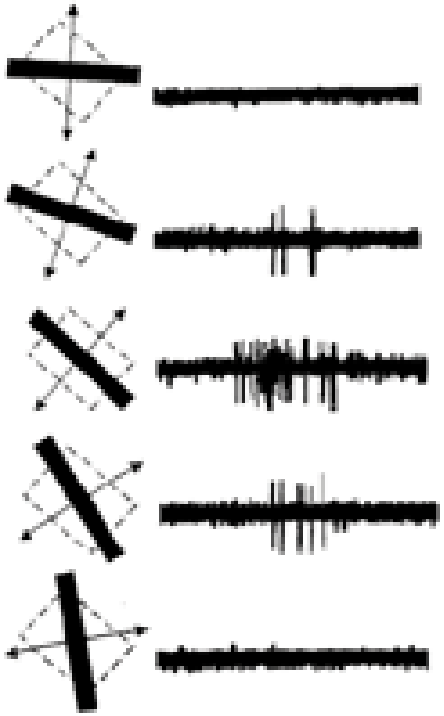
- Gaussian-like tuning operation (and-like)
- Simple units

- Max-like operation (or-like)
- Complex units



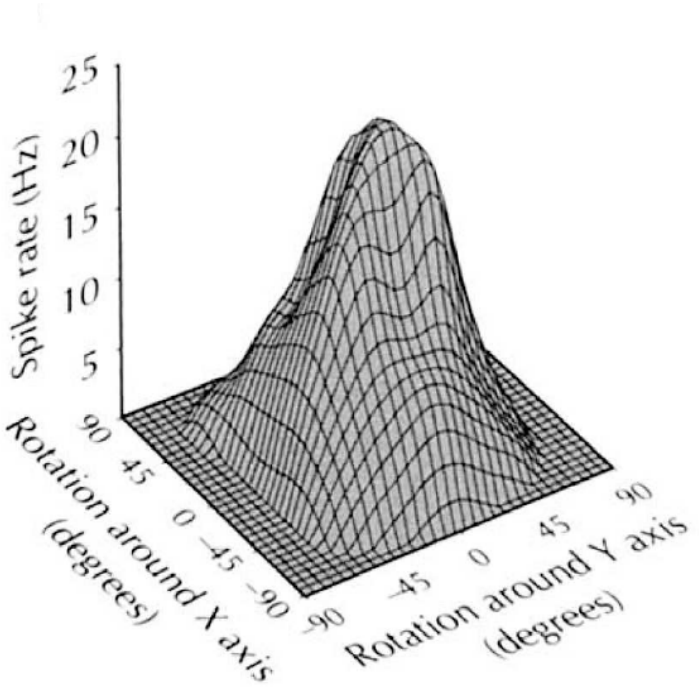
Gaussian tuning

Gaussian tuning in V1 for orientation



Hubel & Wiesel 1958

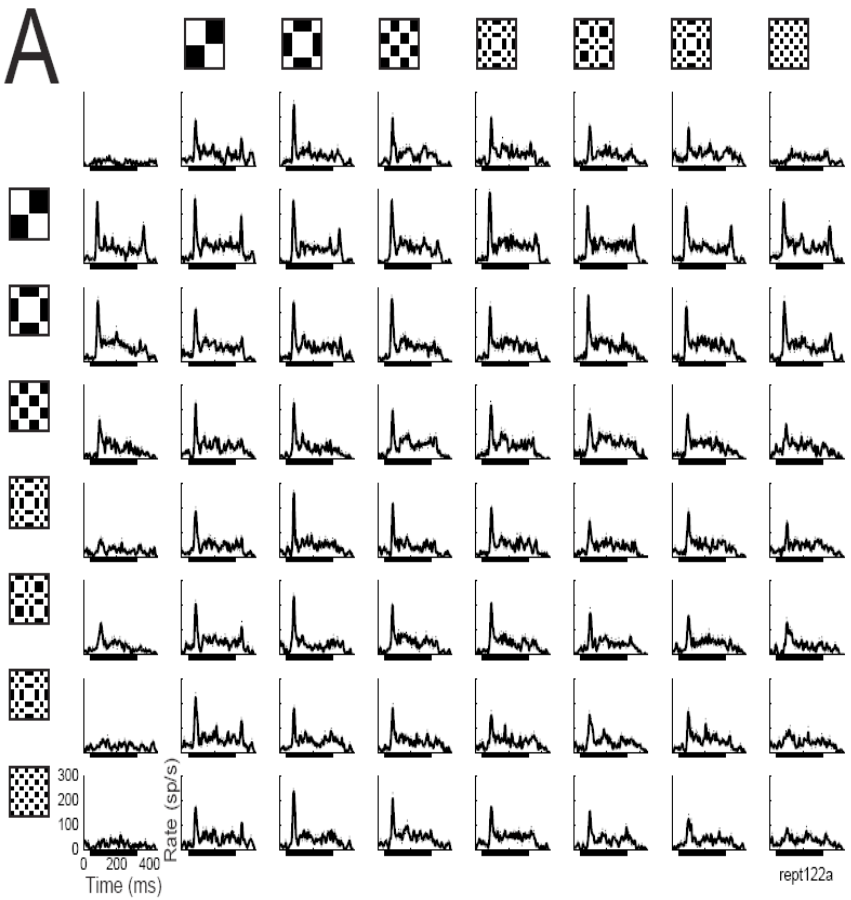
Gaussian tuning in IT around 3D views



Logothetis Pauls & Poggio 1995

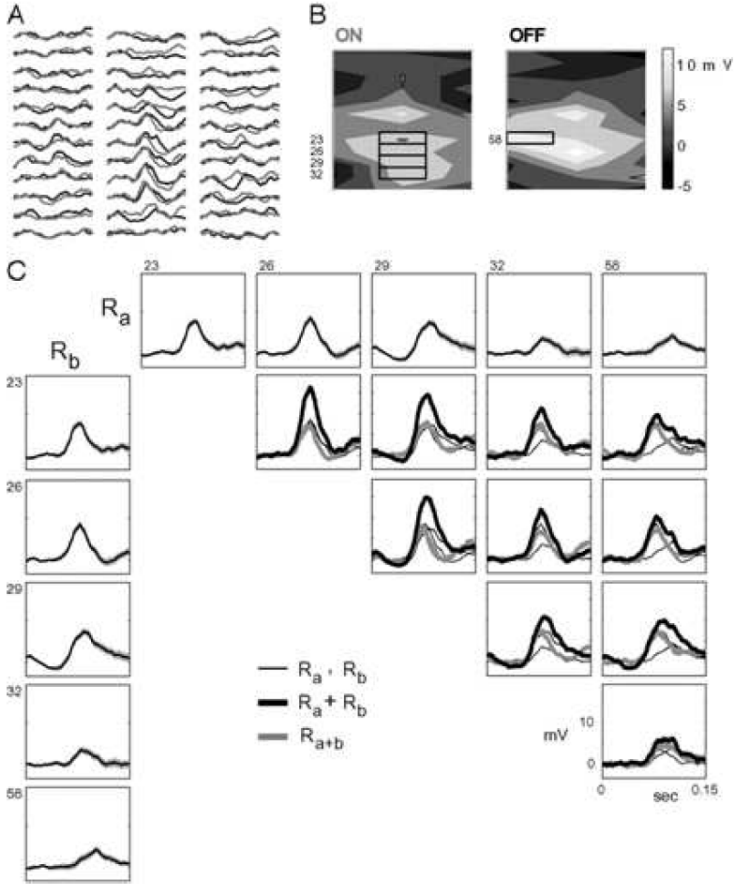
Max-like operation

Max-like behavior in V4



Gawne & Martin 2002

Max-like behavior in V1

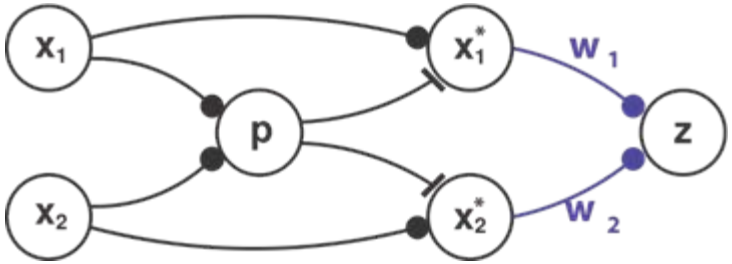


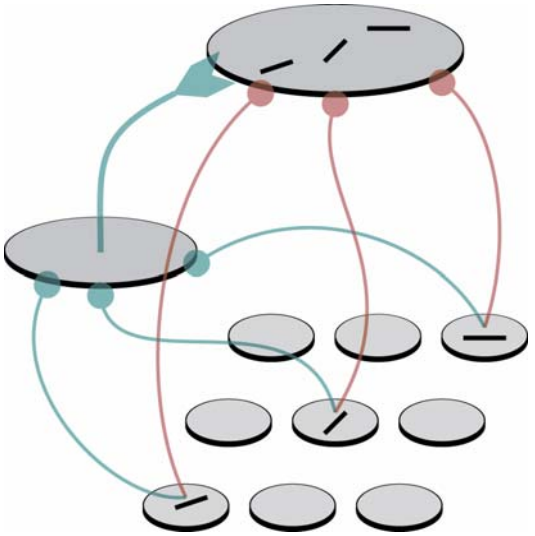
Lampl Ferster Poggio & Riesenhuber 2004
see also Finn Prieber & Ferster 2007

Biophys. implementation

- Max and Gaussian-like tuning can be approximated with same canonical circuit using shunting inhibition

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



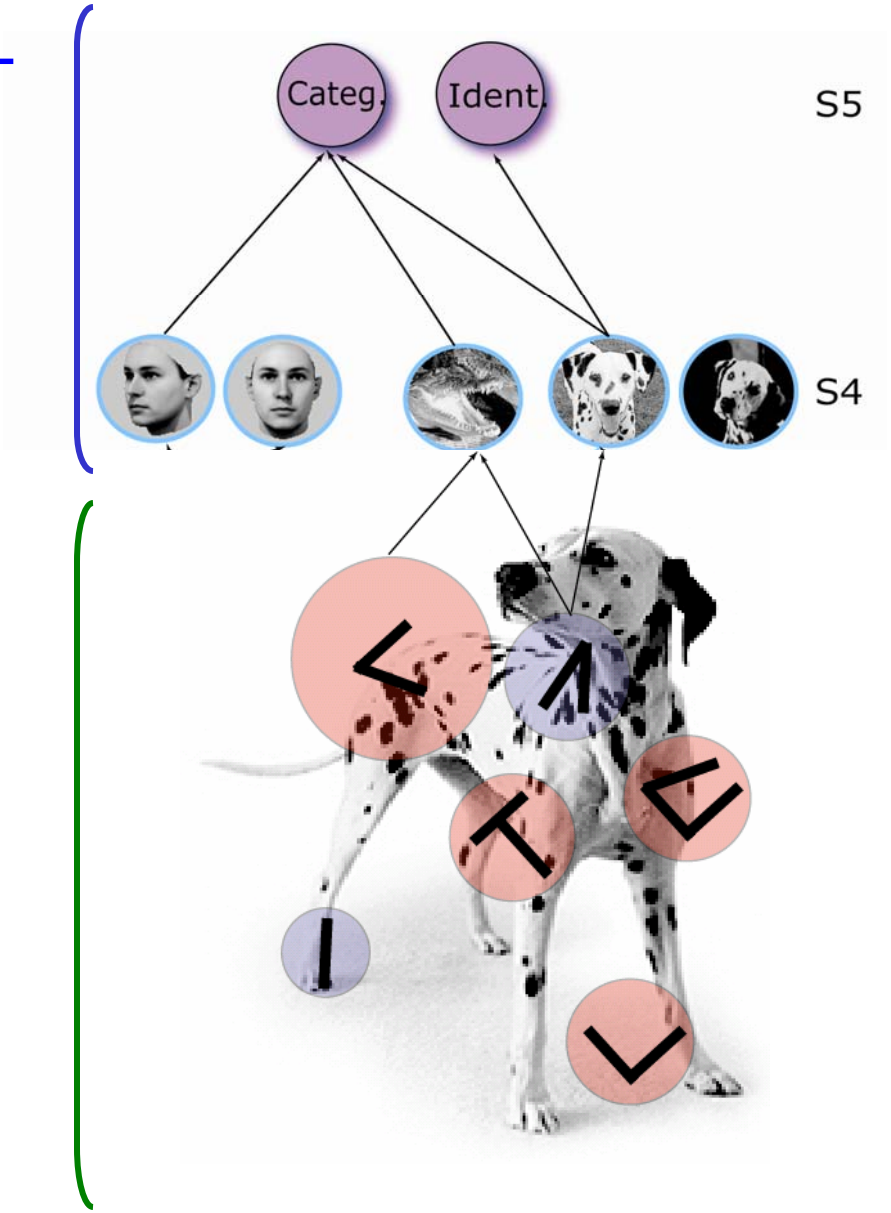


(Knoblich Koch Poggio in prep; Kouh & Poggio 2007; Knoblich Bouvrie Poggio 2007)

Operation	(Steady-State) Output	
Canonical	$y = \frac{\sum_{i=1}^{n} w_i \, {x_i}^p}{k + \left(\sum_{i=1}^{n} {x_i}^q \right)^r}$	Can be implemented by shunting inhibition (Grossberg 1973, Reichardt et al. 1983, Carandini and Heeger, 1994) and spike threshold variability (Anderson et al. 2000, Miller and Troyer, 2002)
Energy Model	$y = \sum_{i=1}^{2} {x_i}^2$	Adelson and Bergen (see also Hassenstein and Reichardt, 1956)
Gaussian-like	$y = \frac{\sum_{i=1}^{n} w_i \, x_i}{k + \sum_{i=1}^{n} {x_i}^2}$	Of the same form as model of MT (Rust et al., Nature Neuroscience, 2007
Max-like	$y = \frac{\sum_{i=1}^{n} {x_i}^3}{k + \sum_{i=1}^{n} {x_i}^2}$	

Task-specific circuits (from IT to PFC)

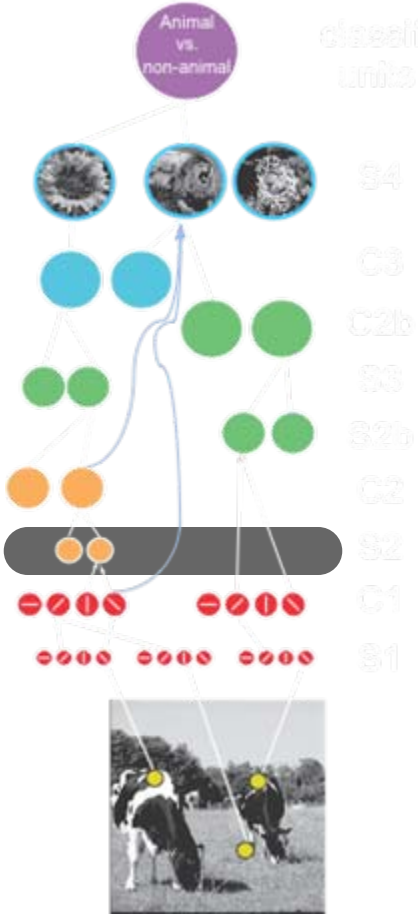
Supervised learning: ~ Gaussian RBF



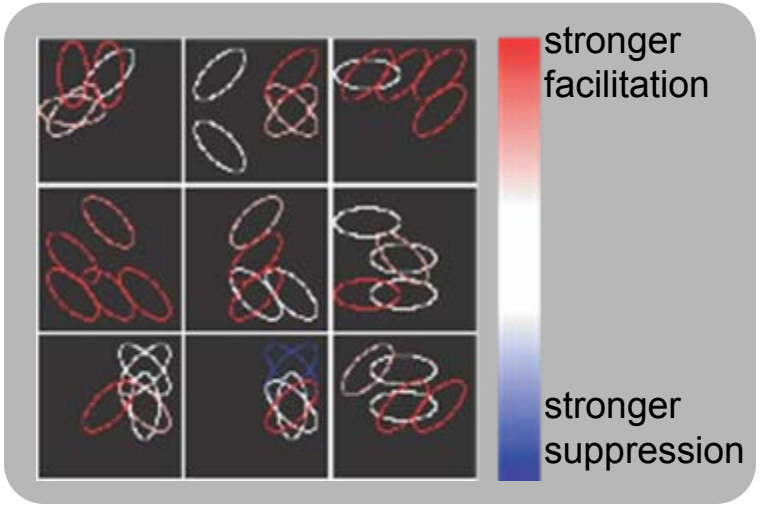
- Generic, overcomplete dictionary of reusable shape components (from V1 to IT) provide unique representation
  - Unsupervised learning (from ~10,000 natural images) during a developmental-like stage

S2 units

- Features of moderate complexity (n~1,000 types)
- Combination of V1-like complex units at different orientations



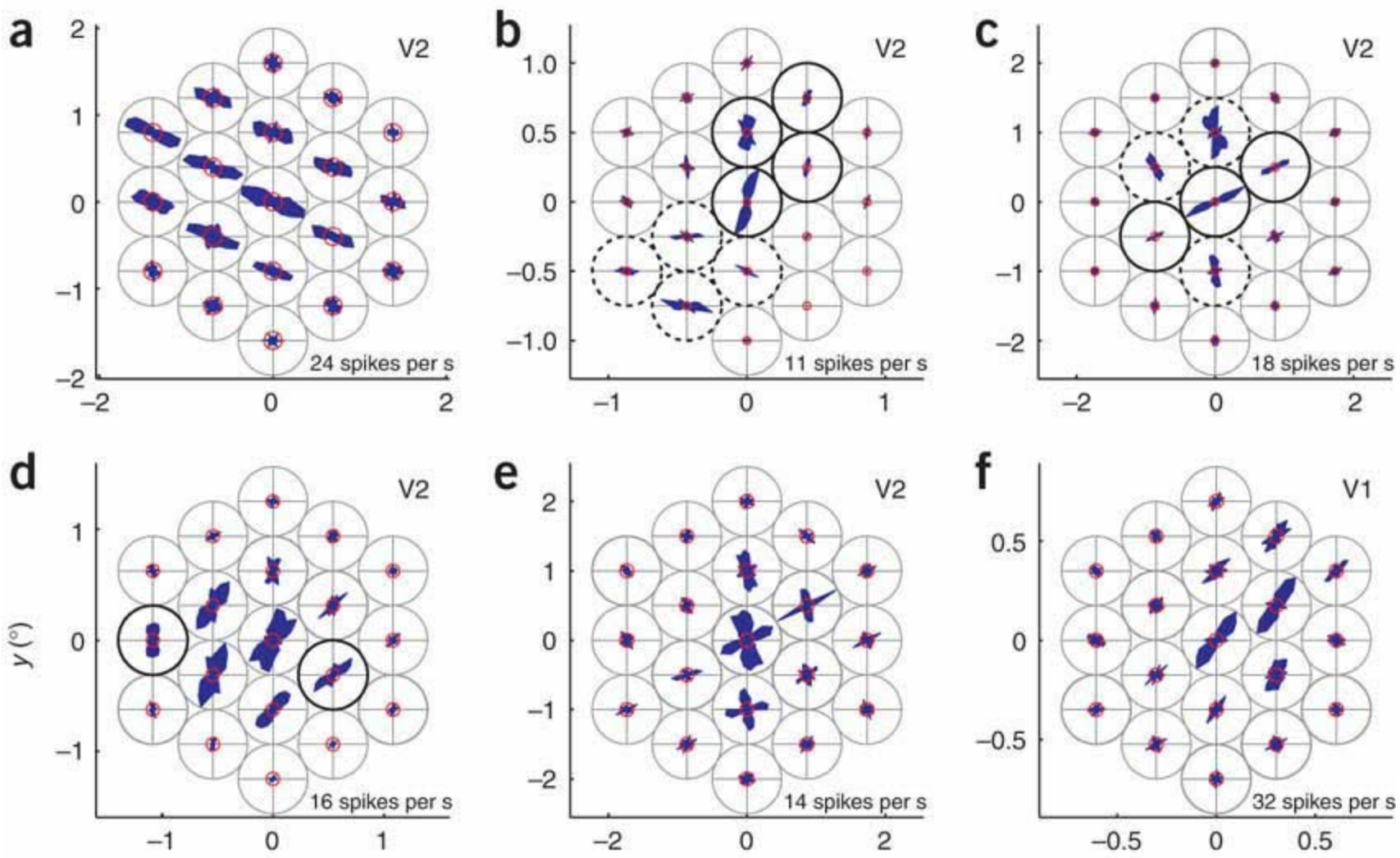
- Synaptic weights ***w*** learned from natural images
- 5-10 subunits chosen at random from all possible afferents (~100-1,000)



Nature Neuroscience - 10, 1313 - 1321 (2007) / Published online: 16 September 2007 | doi:10.1038/nn1975

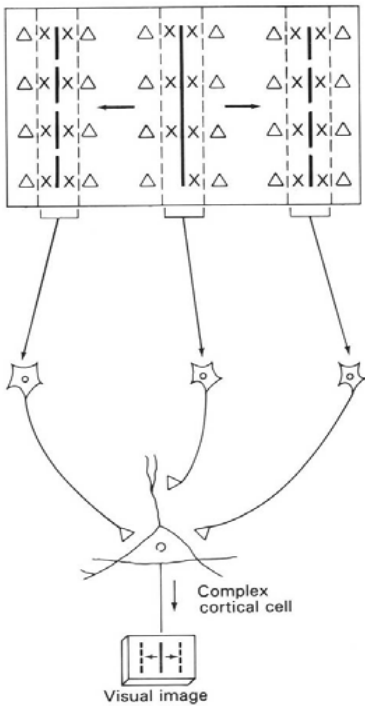
Neurons in monkey visual area V2 encode combinations of orientations

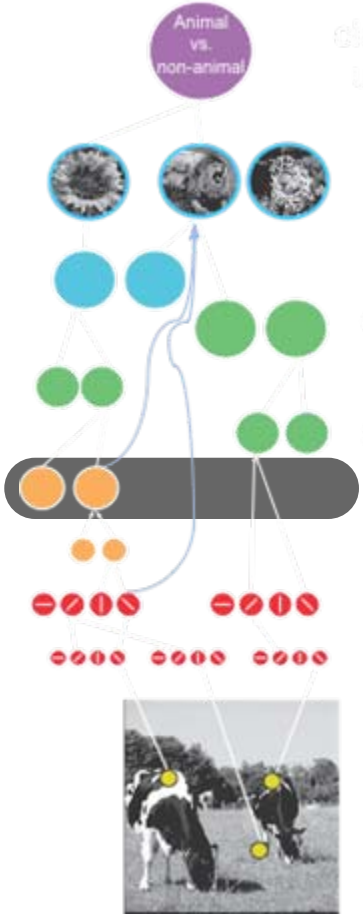
Akiyuki Anzai, Xinmiao Peng & David C Van Essen



C2 units

- Same selectivity as S2 units but increased tolerance to position and size of preferred stimulus
- Local pooling over S2 units with same selectivity but slightly different positions and scales
- A prediction to be tested: S2 units in V2 and C2 units in V4?





A loose hierarchy

- Bypass routes along with main routes:
  - From V2 to TEO (bypassing V4) (Morel & Bullier 1990; Baizer et al 1991; Distler et al 1991; Weller & Steele 1992; Nakamura et al 1993; Buffalo et al 2005)
  - From V4 to TE (bypassing TEO) (Desimone et al 1980; Saleem et al 1992)
- “Replication” of simpler selectivities from lower to higher areas
- Richer dictionary of features with various levels of selectivity and invariance

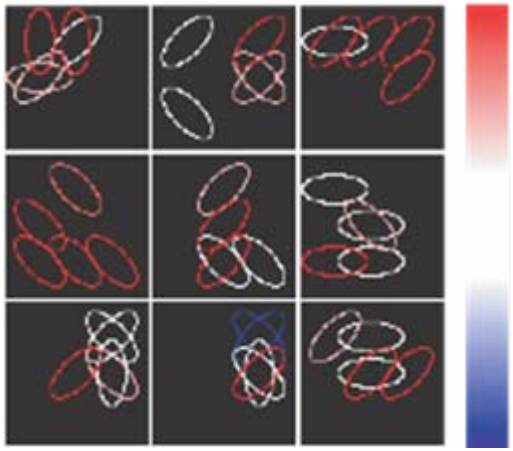
Comparison w| neural data

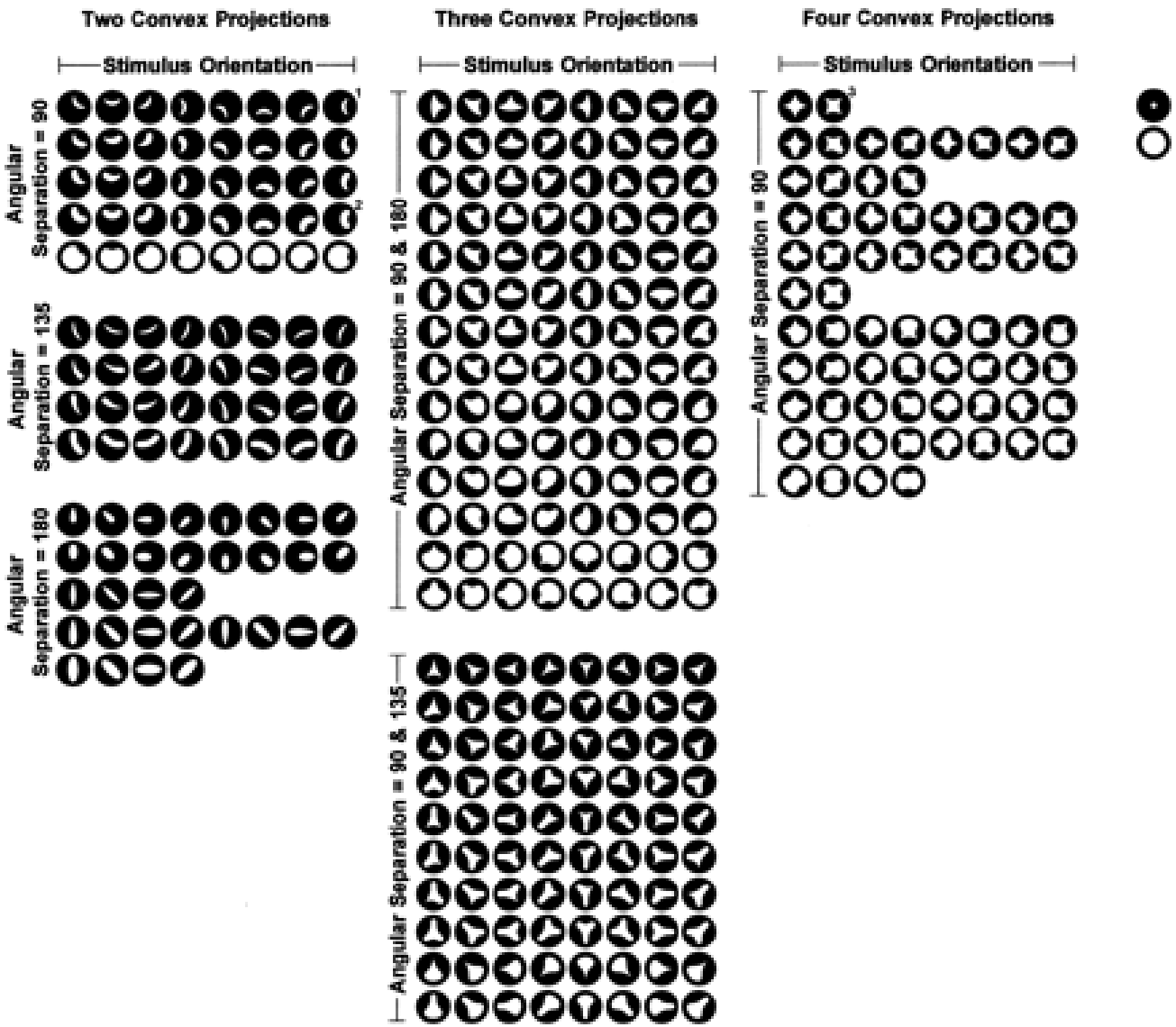
- V1:
  - Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
  - MAX operation in subset of complex cells (Lampl et al 2004)
- V4:
  - Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)
  - MAX operation (Gawne et al 2002)
  - Two-spot interaction (Freiwald et al 2005)
  - Tuning for boundary conformation (Pasupathy & Connor 2001, Cadieu et al., 2007)
  - Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)
- IT:
  - Tuning and invariance properties (Logothetis et al 1995)
  - Differential role of IT and PFC in categorization (Freedman et al 2001, 2002, 2003)
  - Read out data (Hung Kreiman Poggio & DiCarlo 2005)
  - Pseudo-average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo 2007)
- Human:
  - Rapid categorization (Serre Oliva Poggio 2007)
  - Face processing (fMRI + psychophysics) (Riesenhuber et al 2004; Jiang et al 2006)

(Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005)

Comparison w| V4

Tuning for curvature and boundary conformations?

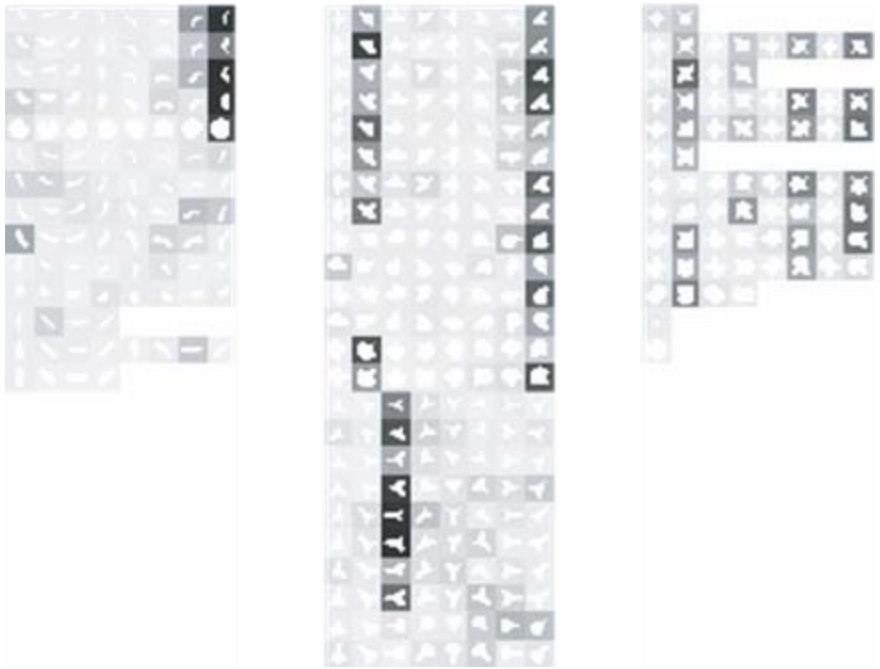




Pasupathy & Connor 2001

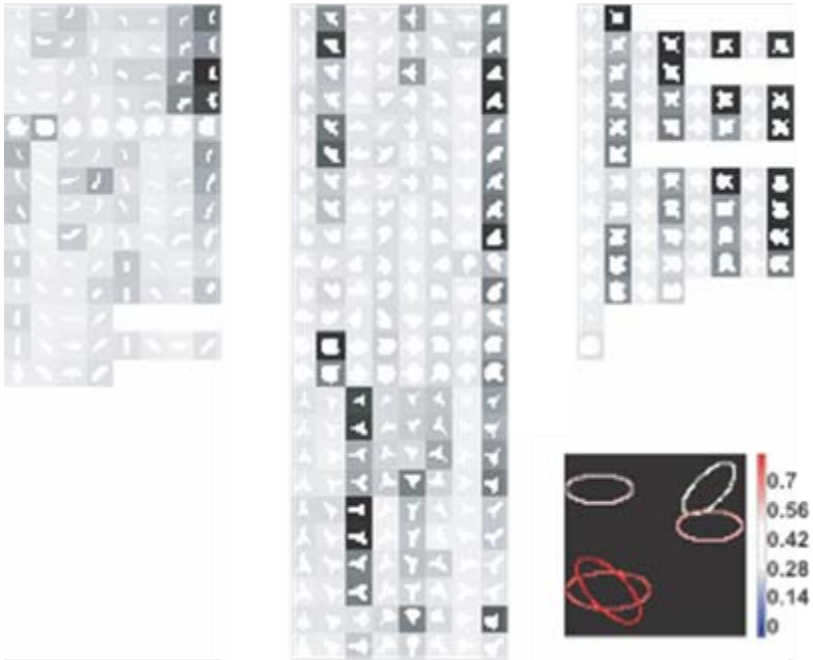
No parameter fitting!

V4 neuron tuned to boundary conformations



Pasupathy & Connor 1999

Most similar model C2 unit



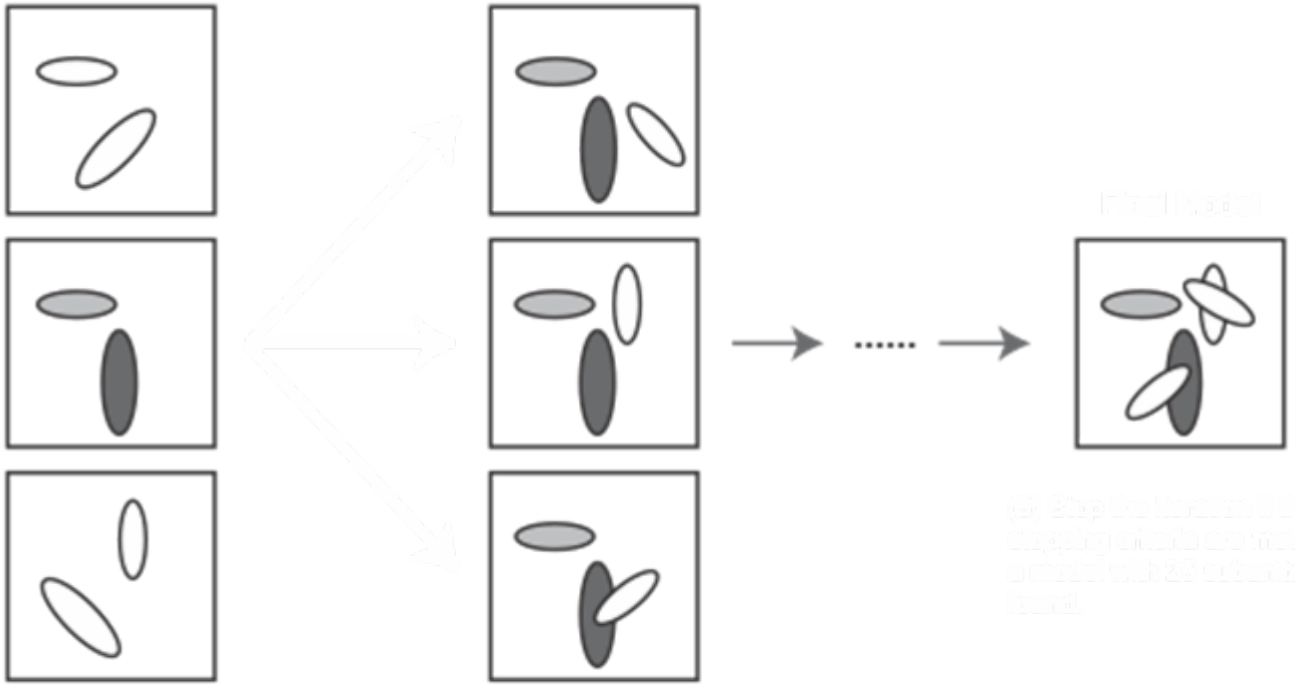
ρ = 0.78

Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005

J Neurophysiol 98: 1733-1750, 2007. First published June 27, 2007

A Model of V4 Shape Selectivity and Invariance

Charles Cadieu, Minjoon Kouh, Anitha Pasupathy, Charles E. Connor, Maximilian Riesenhuber and Tomaso Poggio

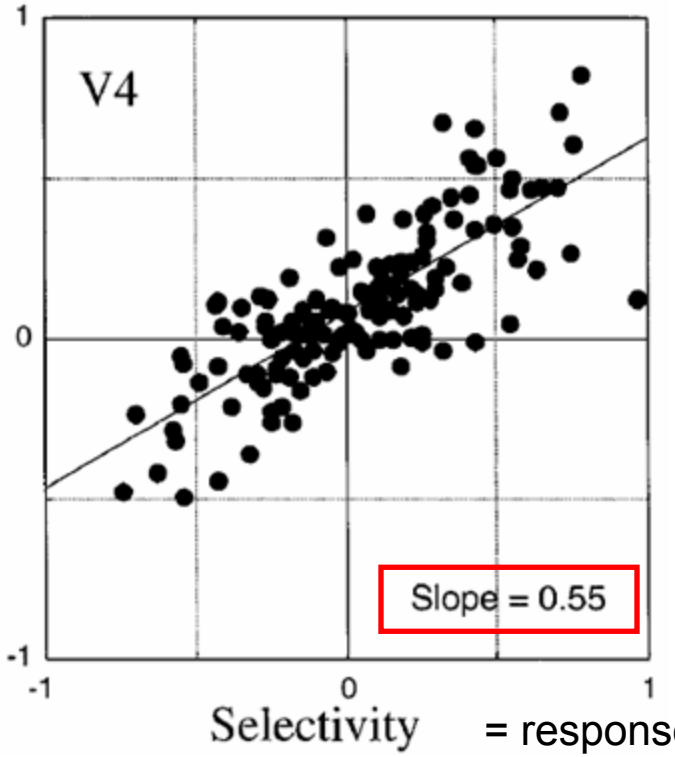


Prediction: Response of the pair is predicted to fall between the responses elicited by the stimuli alone

Sensory Interactions = resp (pair) –resp (reference)

V4 neurons

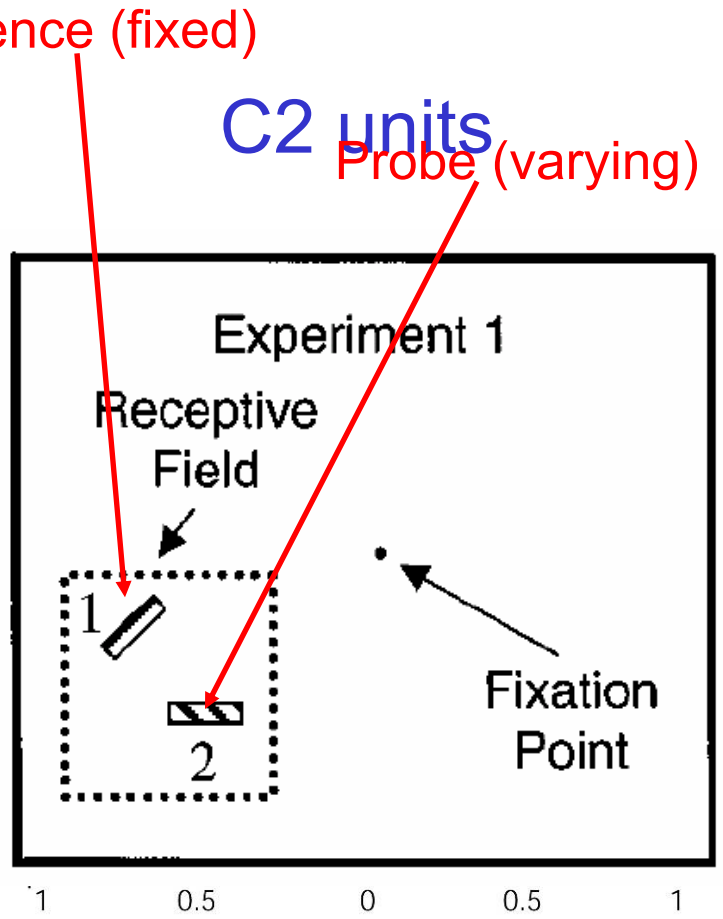
(with attention directed away from receptive field)



= response(probe) –response(reference)

(Reynolds et al 1999)

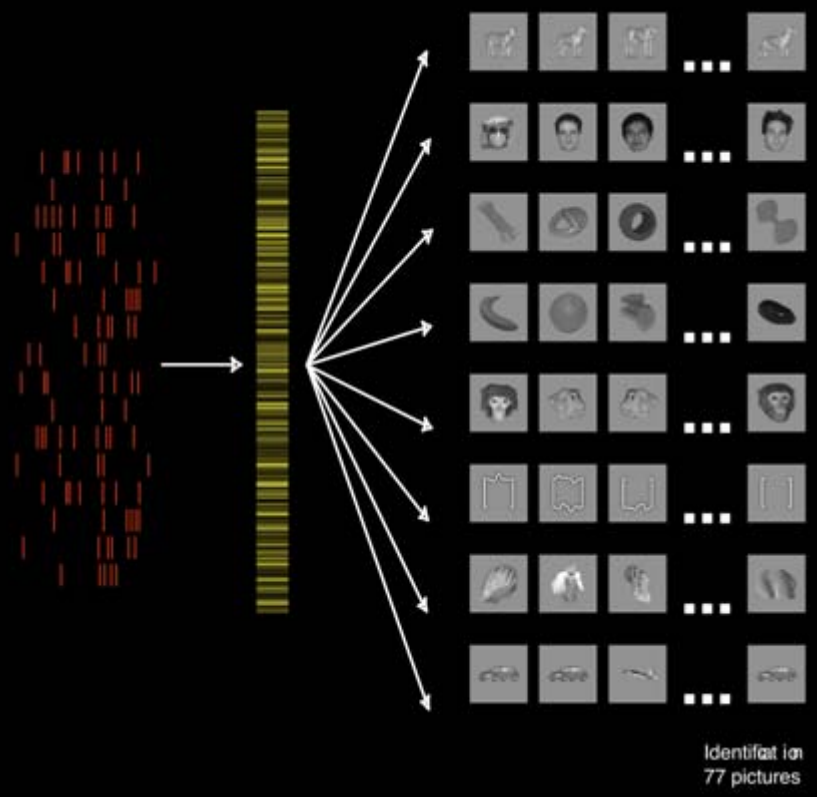
C2 units

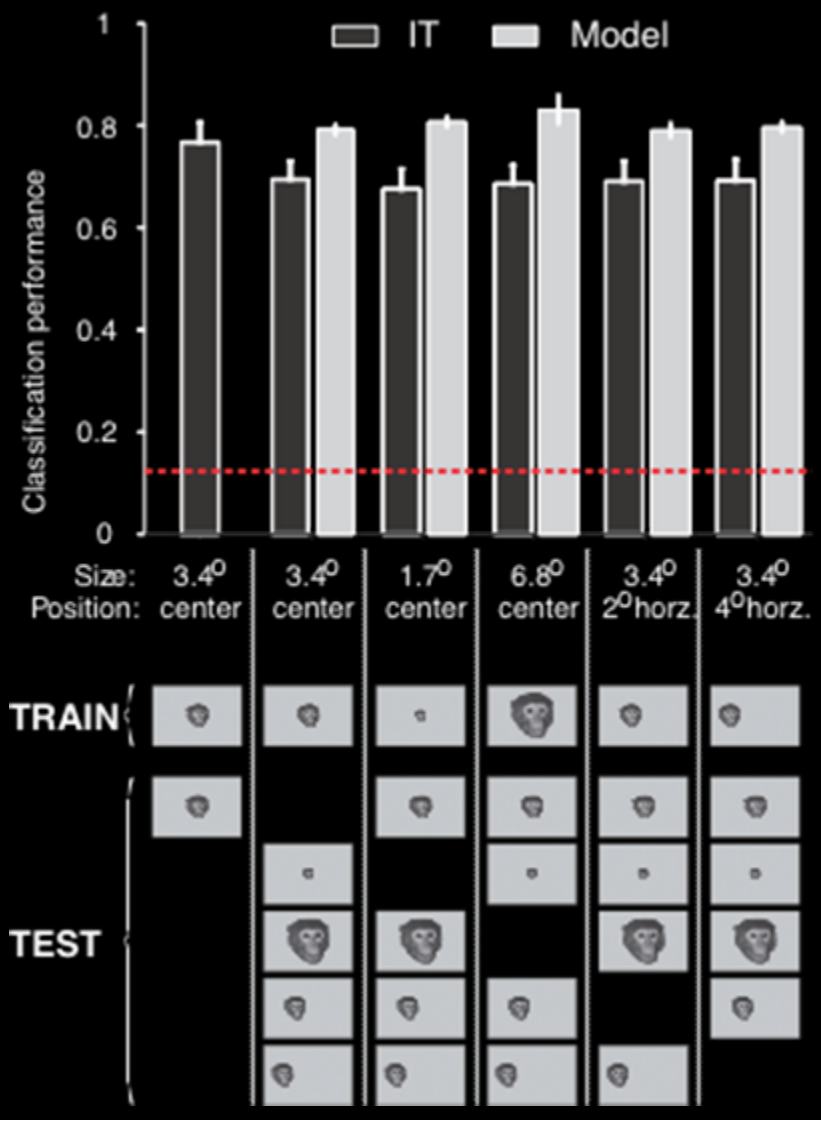


(Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005)

Agreement w| IT Readout data

Hung Kreiman Poggio DiCarlo 2005



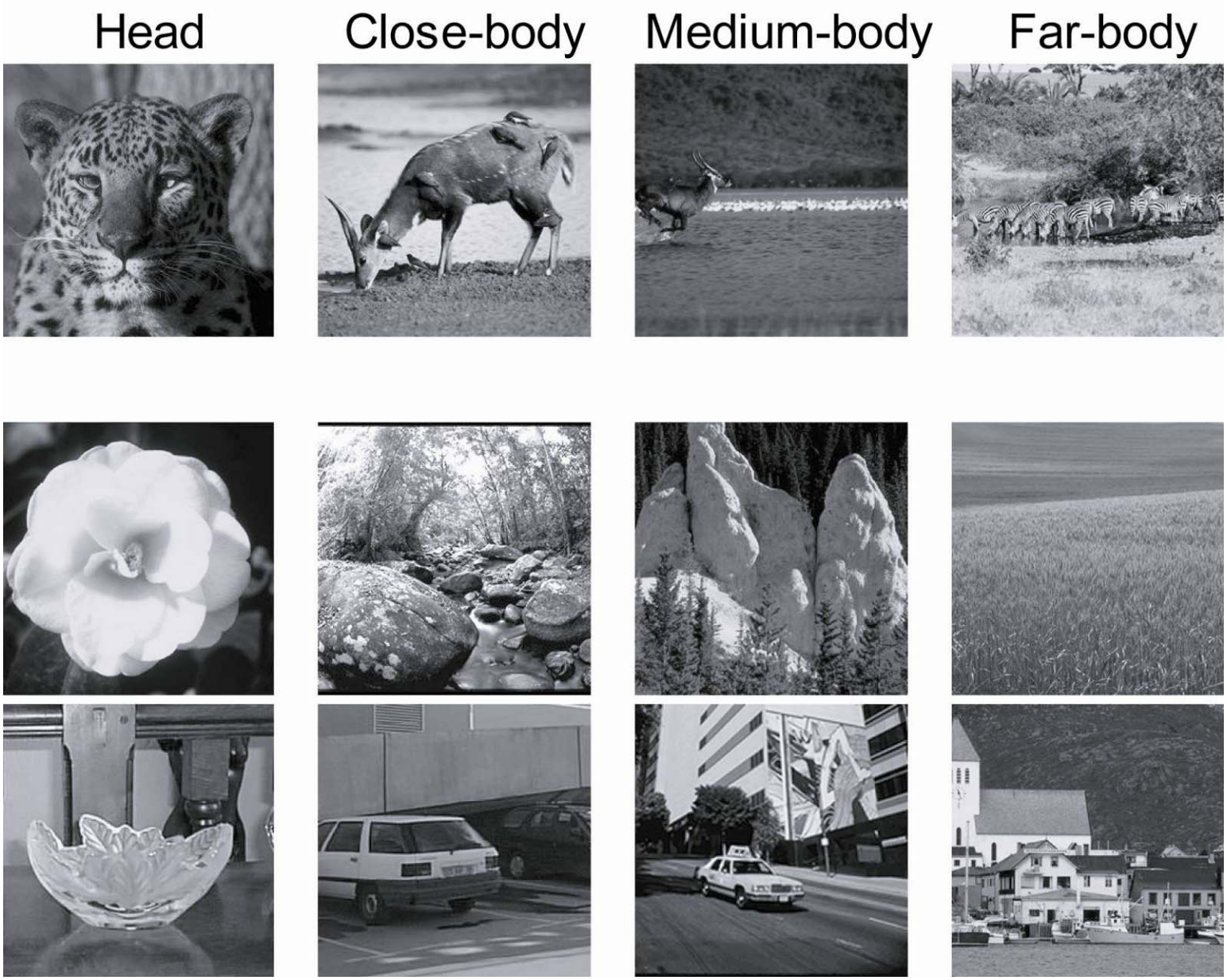


Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005

Remarks

- The stage that includes (V4-PIT)-AIT-PFC represents a learning network of the Gaussian RBF type that is known (from learning theory) to generalize well
- In the theory the stage between IT and ‘’PFC” is a linear classifier – like the one used in the read-out experiments
- The inputs to IT are a large dictionary of selective and invariant features

Rapid categorization

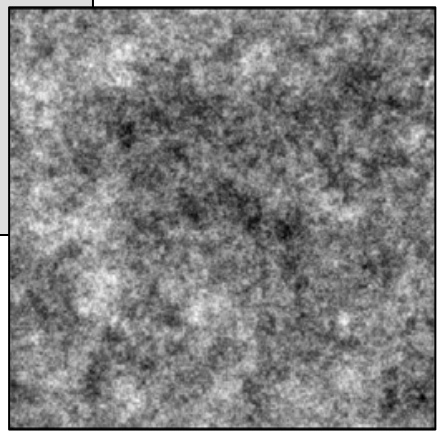


Database collected by Oliva & Torralba

Rapid categorization task (with mask to test feedforward model)

Image

Interval
Image-Mask



Mask
1/f noise

20 ms

30 ms ISI

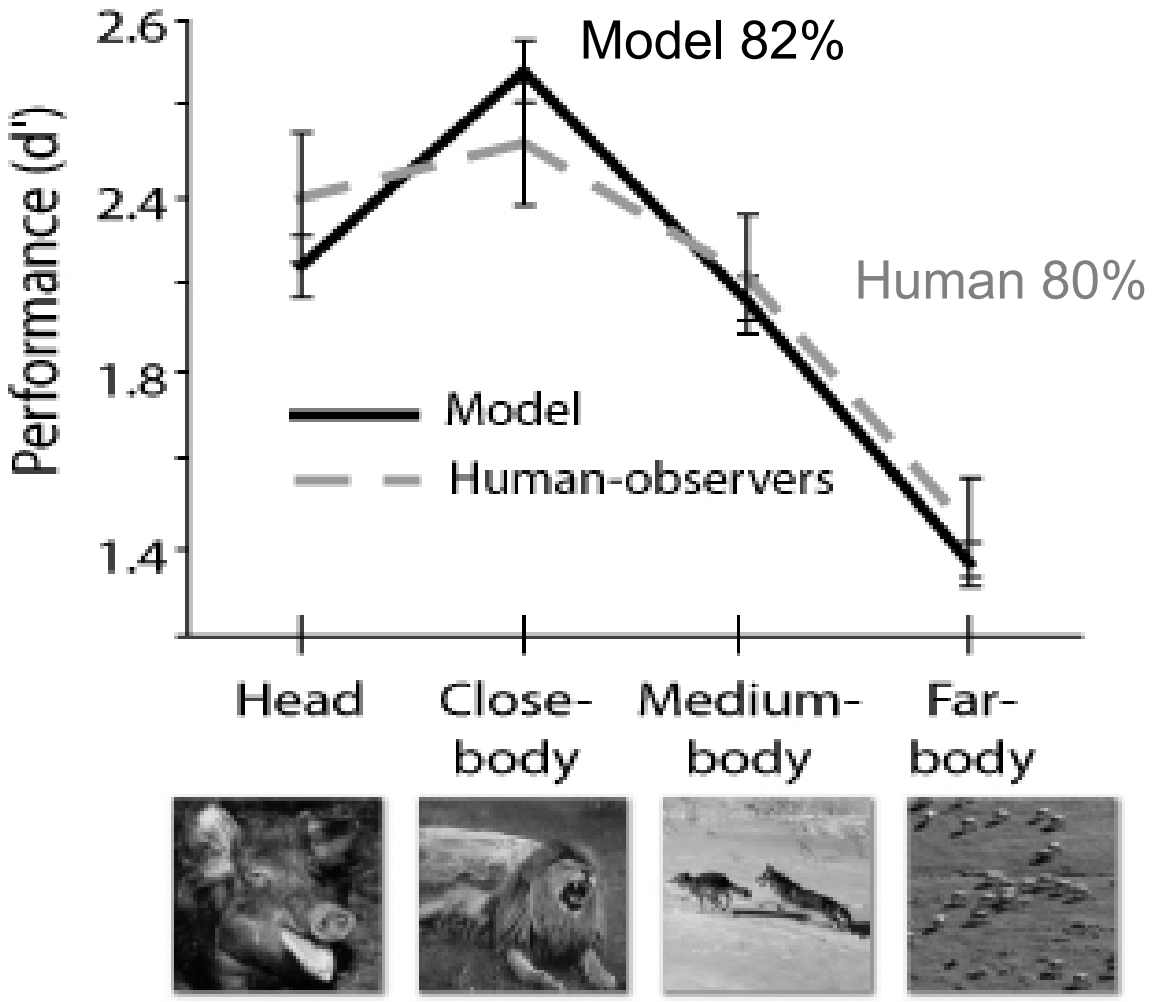
80 ms

Animal present
or not ?

Thorpe et al 1996; Van Rullen & Koch 2003; Bacon-Mace et al 2005

...solves the problem (when mask forces feedforward processing)...

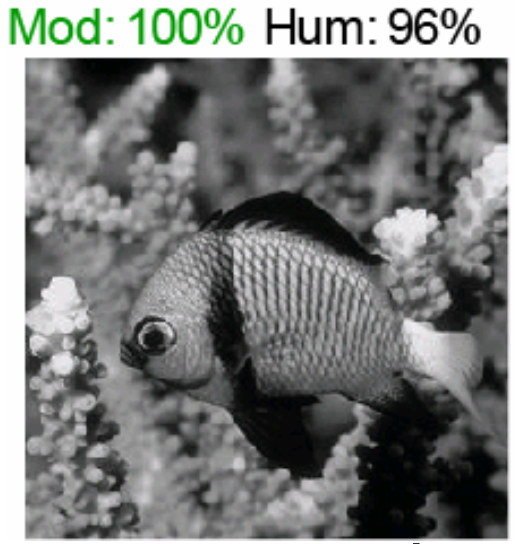
- d'~ standardized error rate
- the higher the d', the better the perf.



Serre Oliva & Poggio 2007

Further comparisons

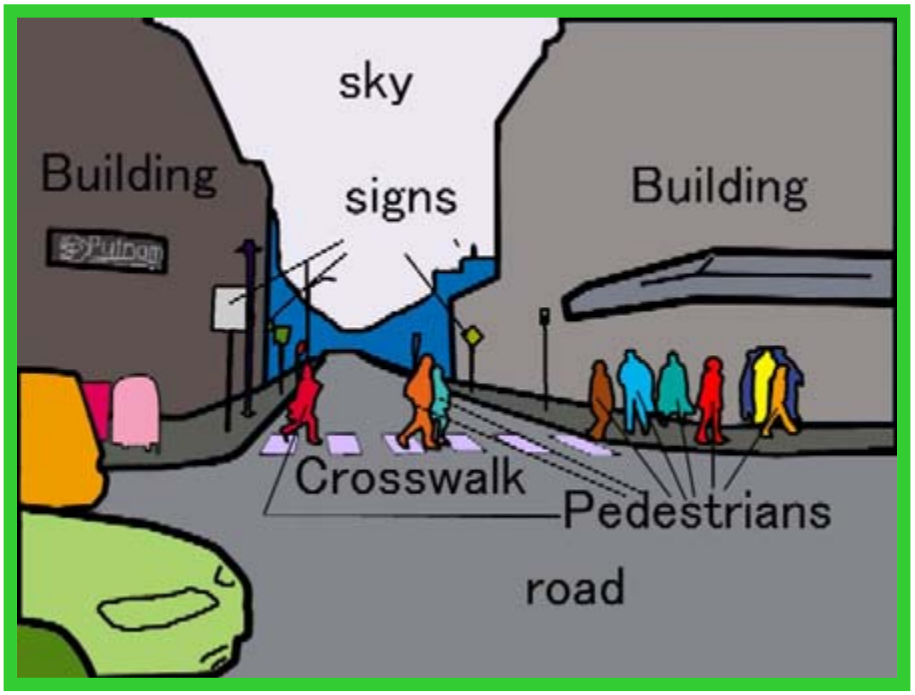
- Image-by-image correlation:
  - Heads: ρ=0.71
  - Close-body: ρ=0.84
  - Medium-body: ρ=0.71
  - Far-body: ρ=0.60



- Model predicts level of performance on rotated images (90 deg and inversion)

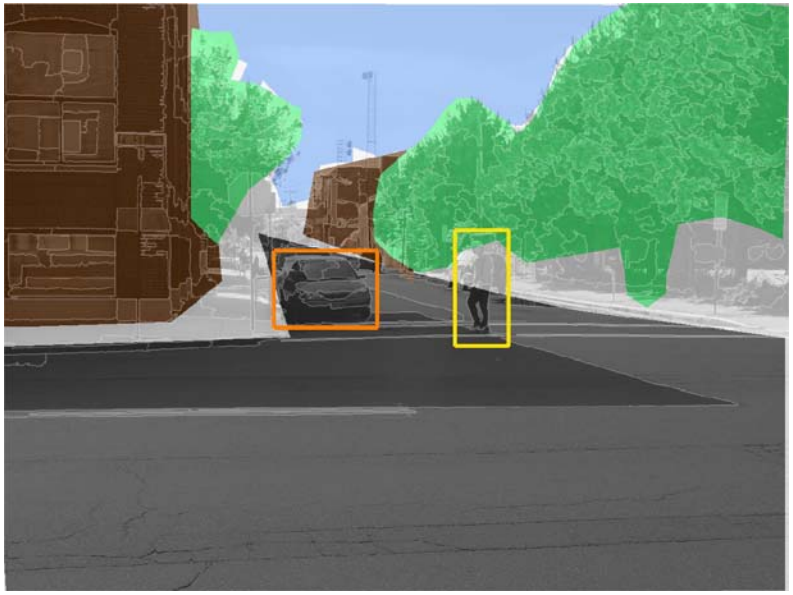
Serre Oliva & Poggio PNAS 2007

The street scene project



Source: Bileschi & Wolf

The StreetScenes Database

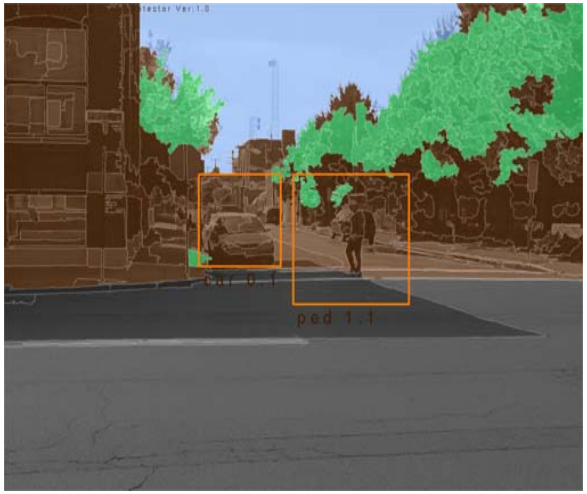


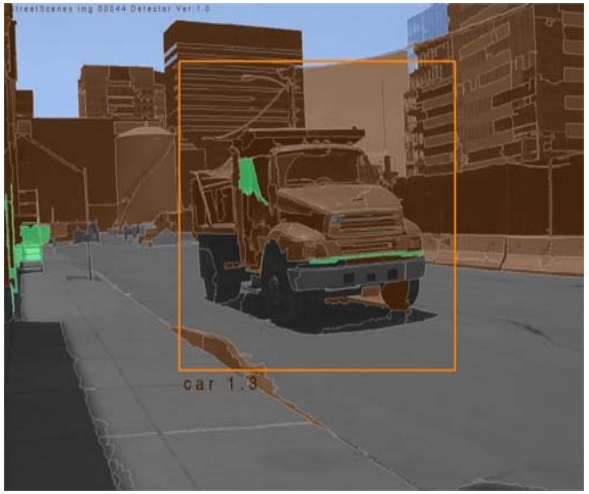
3,547 Images, all taken with the same camera, of the same type of scene, and hand labeled with the same objects, using the same labeling rules.

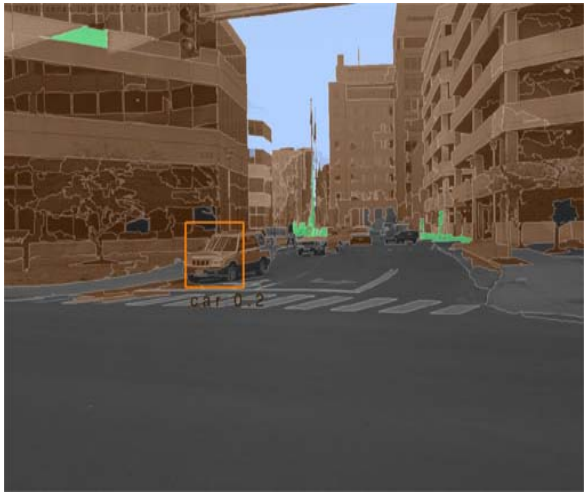
Object	car	pedestrian	bicycle	building	tree	road	sky
# Labeled Examples	5799	1449	209	5067	4932	3400	2562

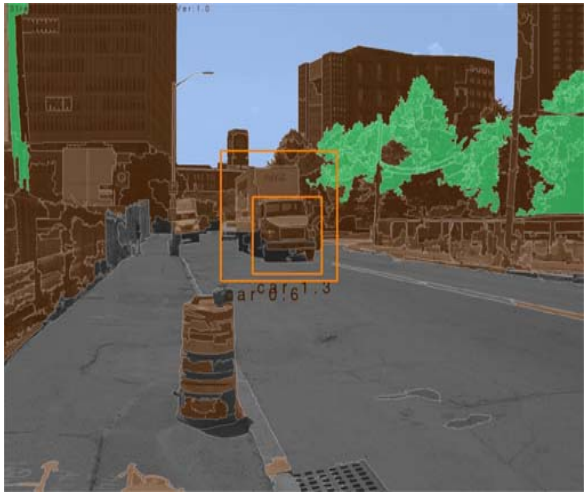
http://cbcl.mit.edu/software-datasets/streetscenes/

Examples

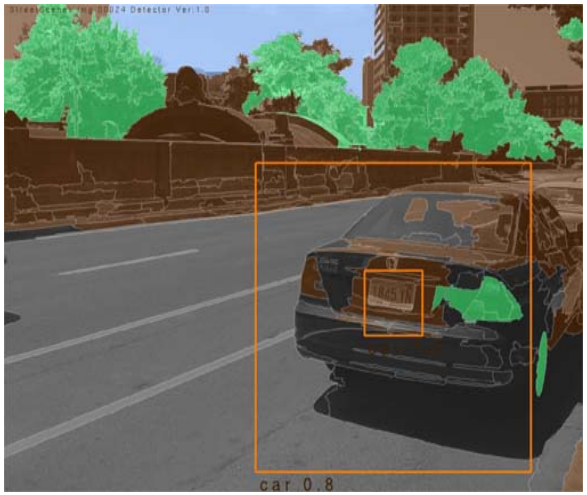


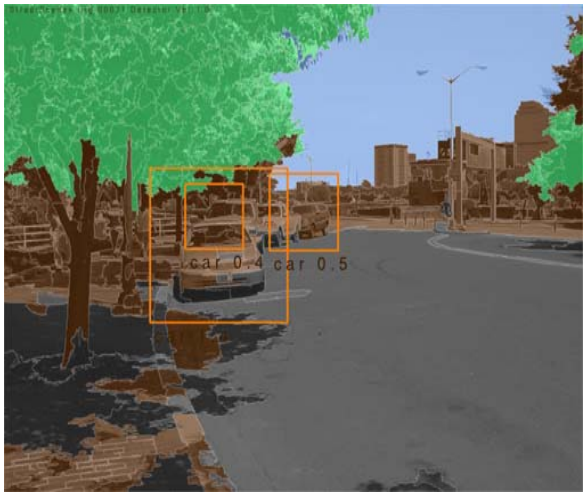


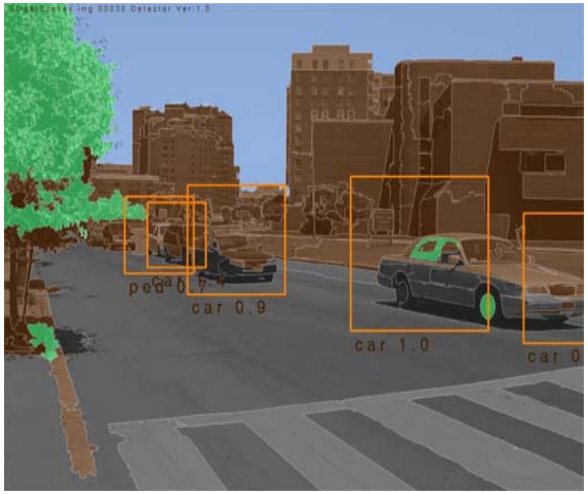


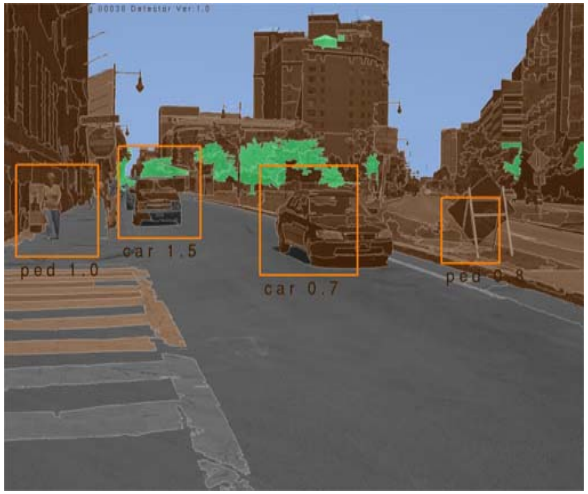


Examples

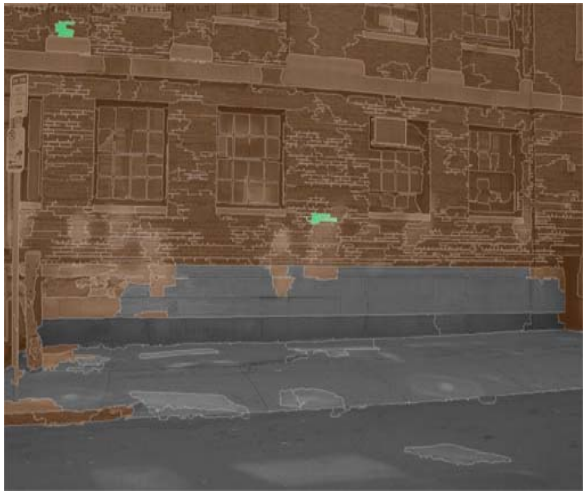


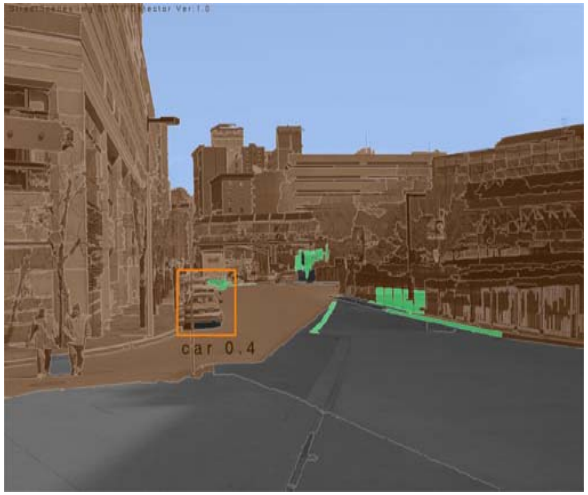




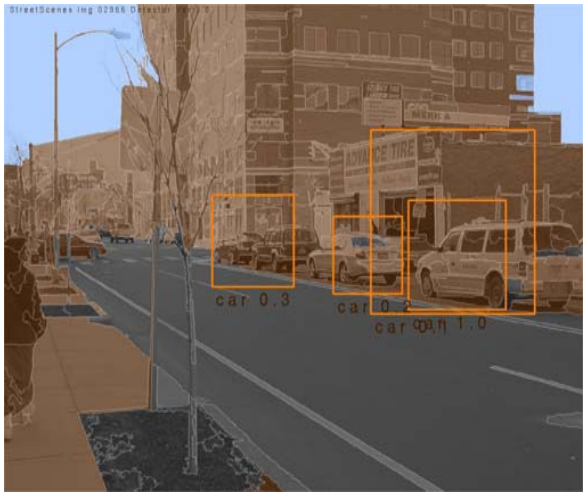


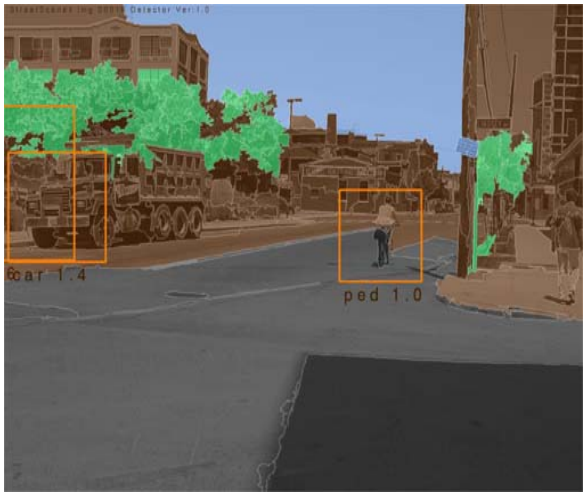
Examples

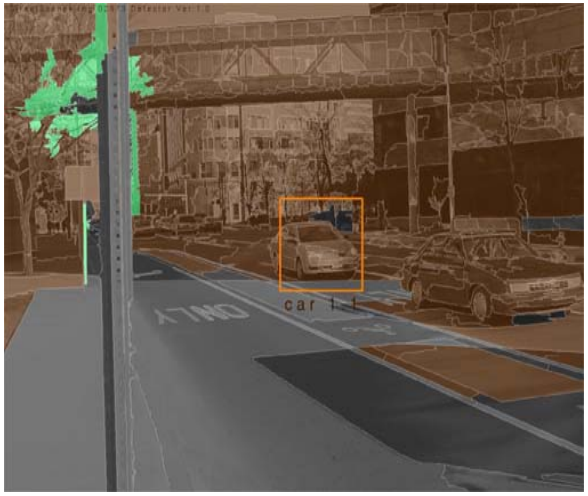


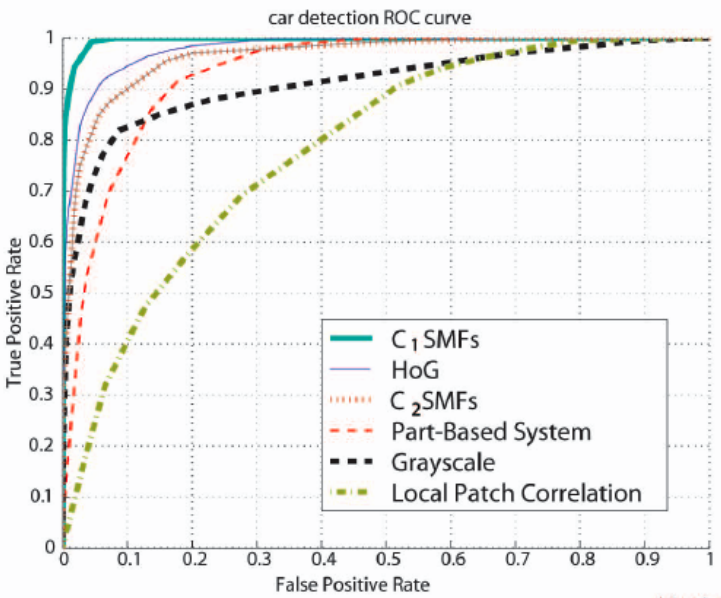


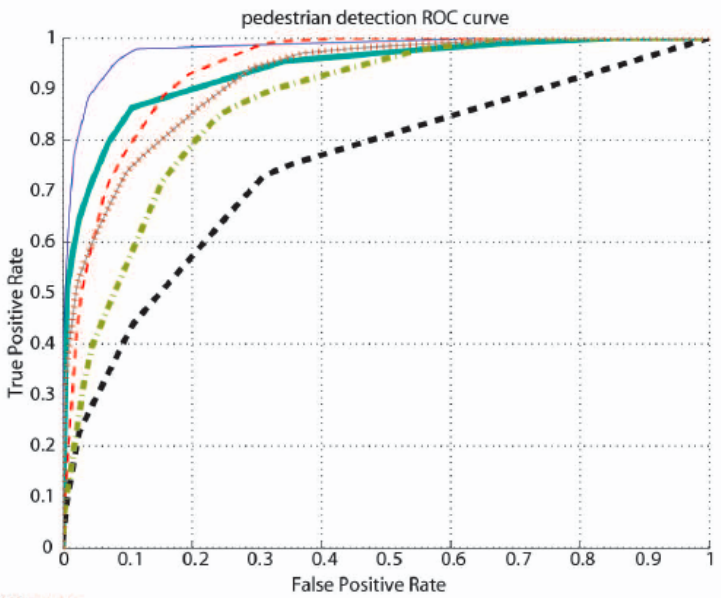
Examples

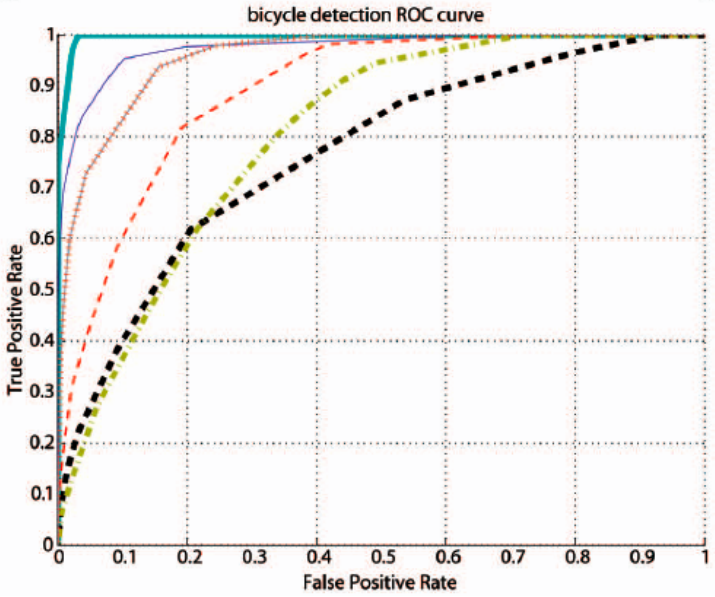






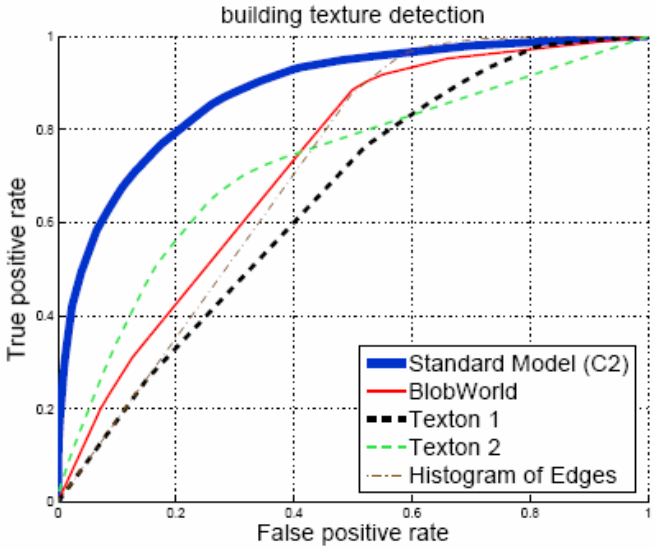


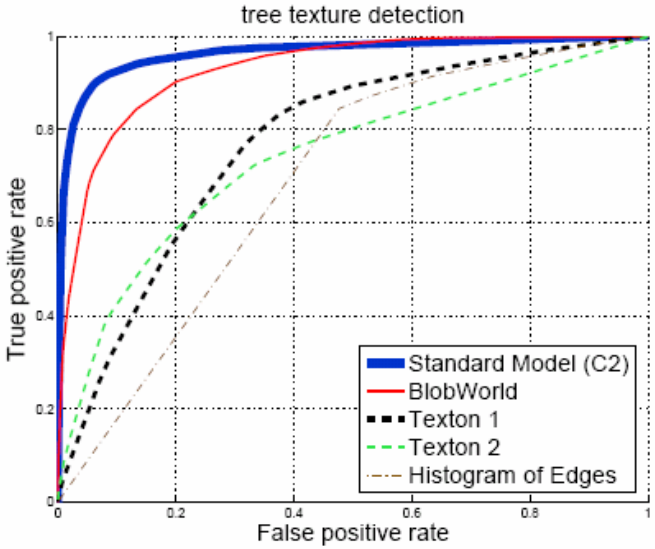


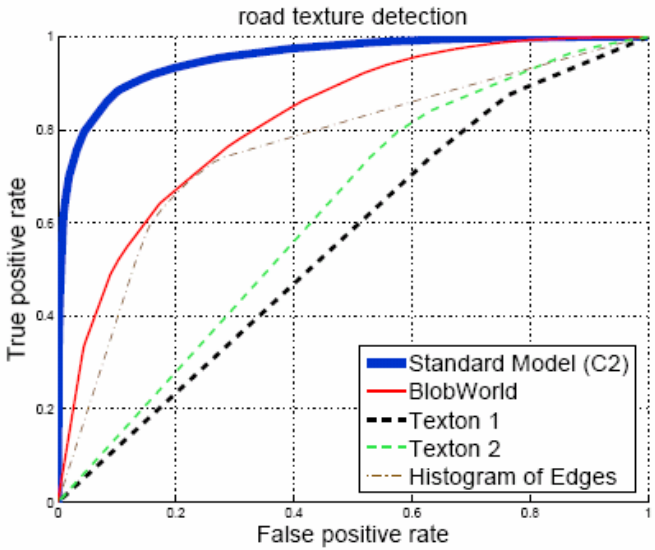


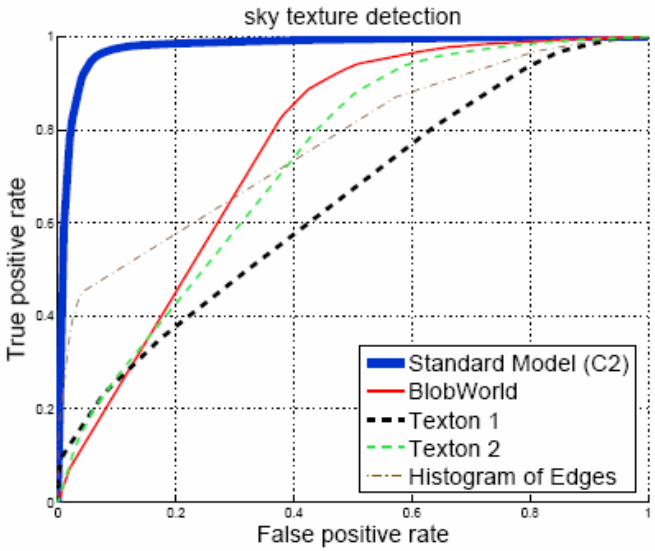
- HoG:
(Dalal & Triggs 2005)
- Part-based system:
(Leibe et al 2004)
- Local patch correlation:
(Torralba et al 2004)

Serre Wolf Bileschi Riesenhuber & Poggio PAMI 2007









Serre Wolf Bileschi Riesenhuber & Poggio PAMI 2007

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

What is next

- A challenge for physiology: disprove basic aspects of the architecture
- Extensions to color and stereo
- More sophisticated unsupervised, developmental learning in V1, V4, PIT: how?
- Extension to time and videos
- Extending the simulation to integrate-and-fire neurons (~ 1 billion) and realistic synapses: towards the neural code

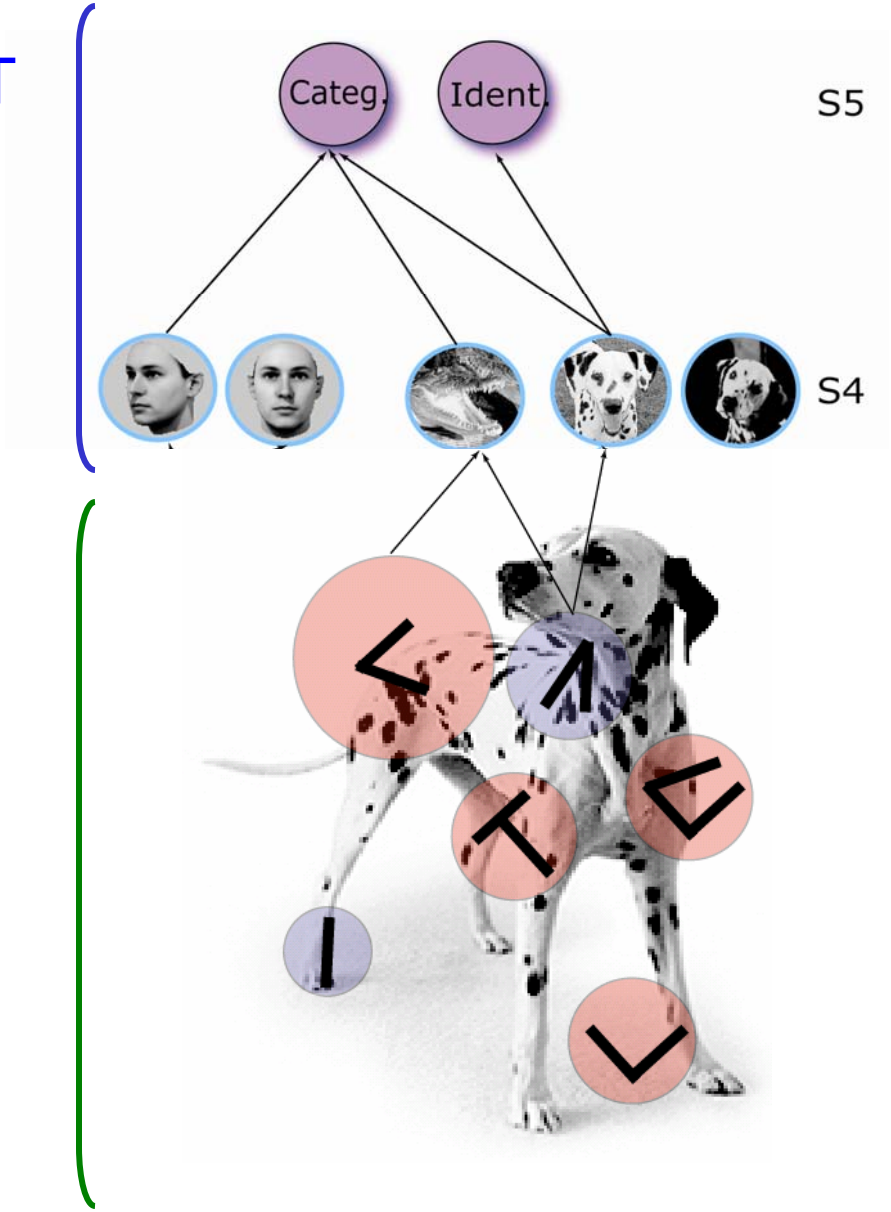
What is next

- A challenge for physiology: disprove basic aspects of the architecture
- Extensions to color and stereo
- More sophisticated unsupervised, developmental learning in V1, V4, PIT: how?
- Extension to time and videos
- Extending the simulation to integrate-and-fire neurons (~ 1 billion) and realistic synapses: towards the neural code

Layers of cortical processing units

Task-specific circuits (from IT to PFC)

Supervised learning



- Generic dictionary of shape components (from V1 to IT)
  - Unsupervised learning during a developmental-like stage learning dictionaries of “templates” at different S levels

Learning the invariance from temporal continuity

w| T. Masquelier & S. Thorpe (CNRS, France)

✦ Simple cells learn correlation in space (at the same time)

✦ Complex cells learn correlation in time

SHOW MOVIE

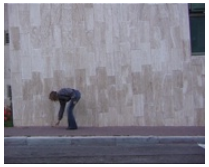
Foldiak 1991; Perrett et al 1984; Wallis & Rolls, 1997; Einhauser et al 2002; Wiskott & Sejnowski 2002; Spratling 2005

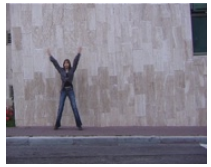
What is next

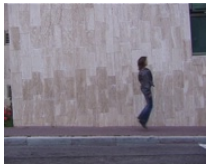
- A challenge for physiology: disprove basic aspects of the architecture
- Extensions to color and stereo
- More sophisticated unsupervised, developmental learning in V1, V4, PIT: how?
- Extension to time and videos
- Extending the simulation to integrate-and-fire neurons (~ 1 billion) and realistic synapses: towards the neural code

The problem

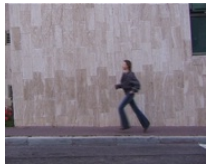
Training Videos

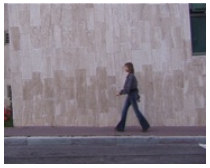

bend

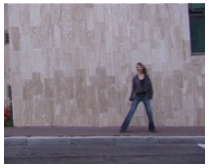

jack

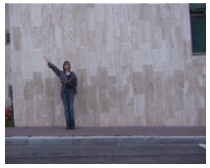

jump

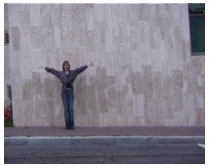

pjump


run


walk


side


wave1

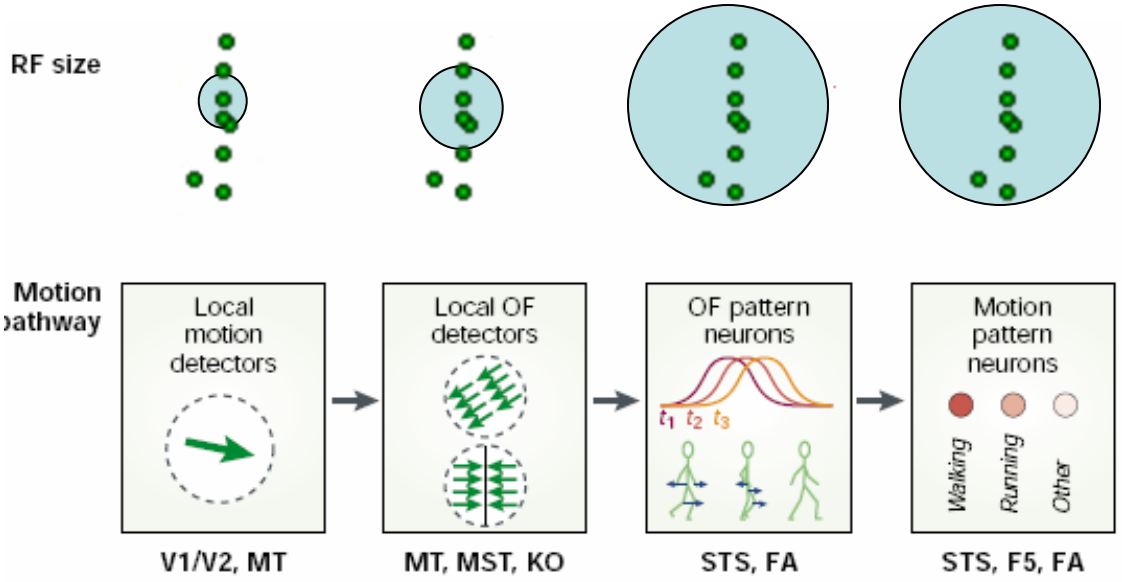

wave2

Testing videos

*each video~4s, 50~100 frames

Dataset from (Blank et al, 2005)

Previous work:
recognizing biological motion using a model of the dorsal stream

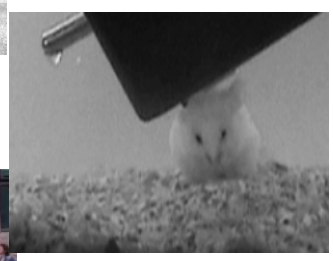


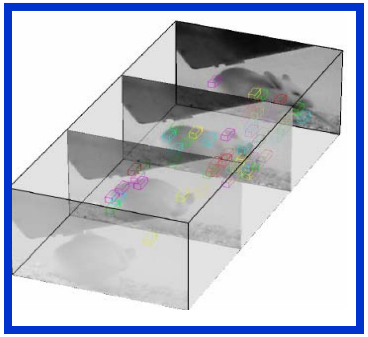
Adapted from (Giese & Poggio, 2003)

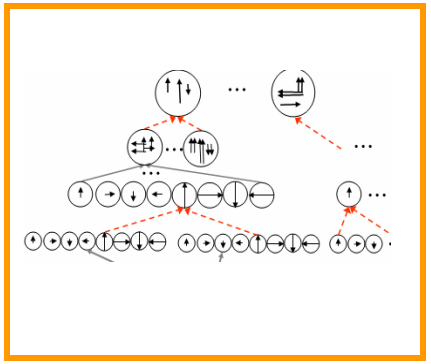
See also (Casile & Giese 2005; Sigala et al, 2005)

Multi-class recognition accuracy

	Baseline	Our system
KTH Human	81.3 %	**91.6 %**
UCSD Mice	75.6 %	**79.0 %**
Weiz. Human	86.7 %	**96.3 %**
Average	**81.2 %**	**89.6 %**





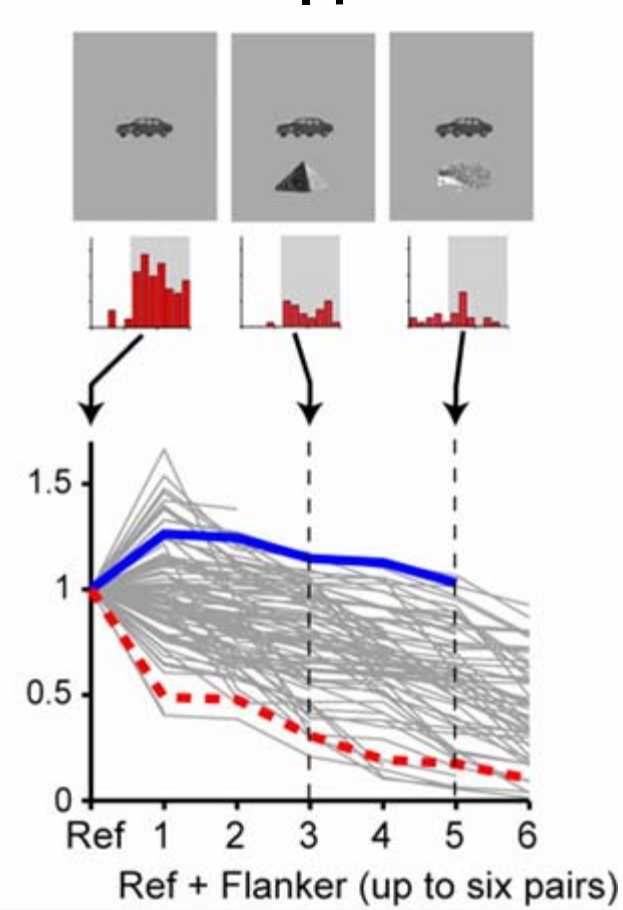


HH. Jhuang, T. Serre, L. Wolf* and T. Poggio, ICCV, 2007

* chances: 10%~20%

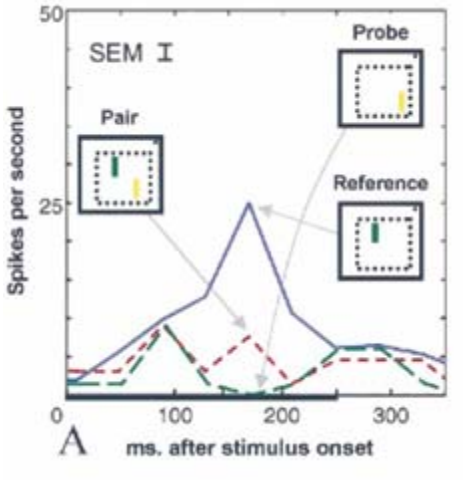
What is next:
beyond feedforward models: limitations

IT



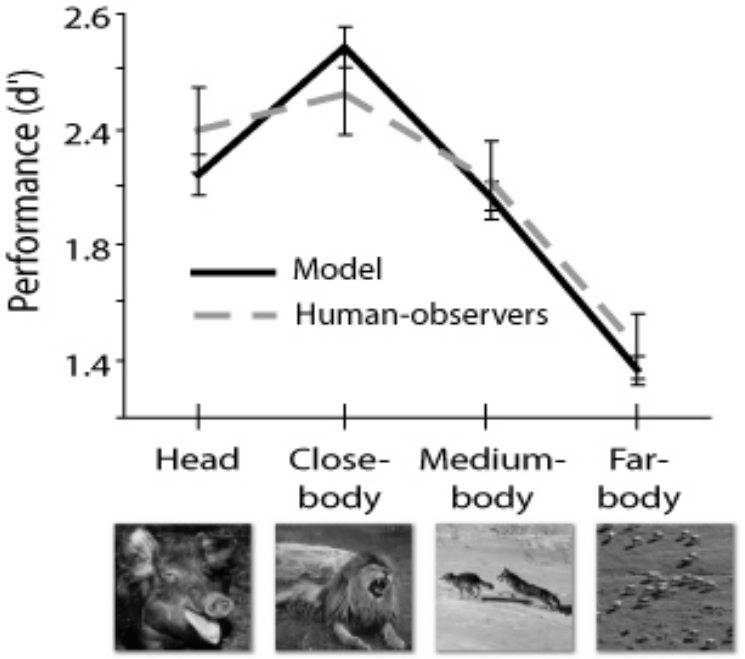
Zoccolan Kouh Poggio DiCarlo 2007

V4



Reynolds Chelazzi & Desimone 1999

Psychophysics

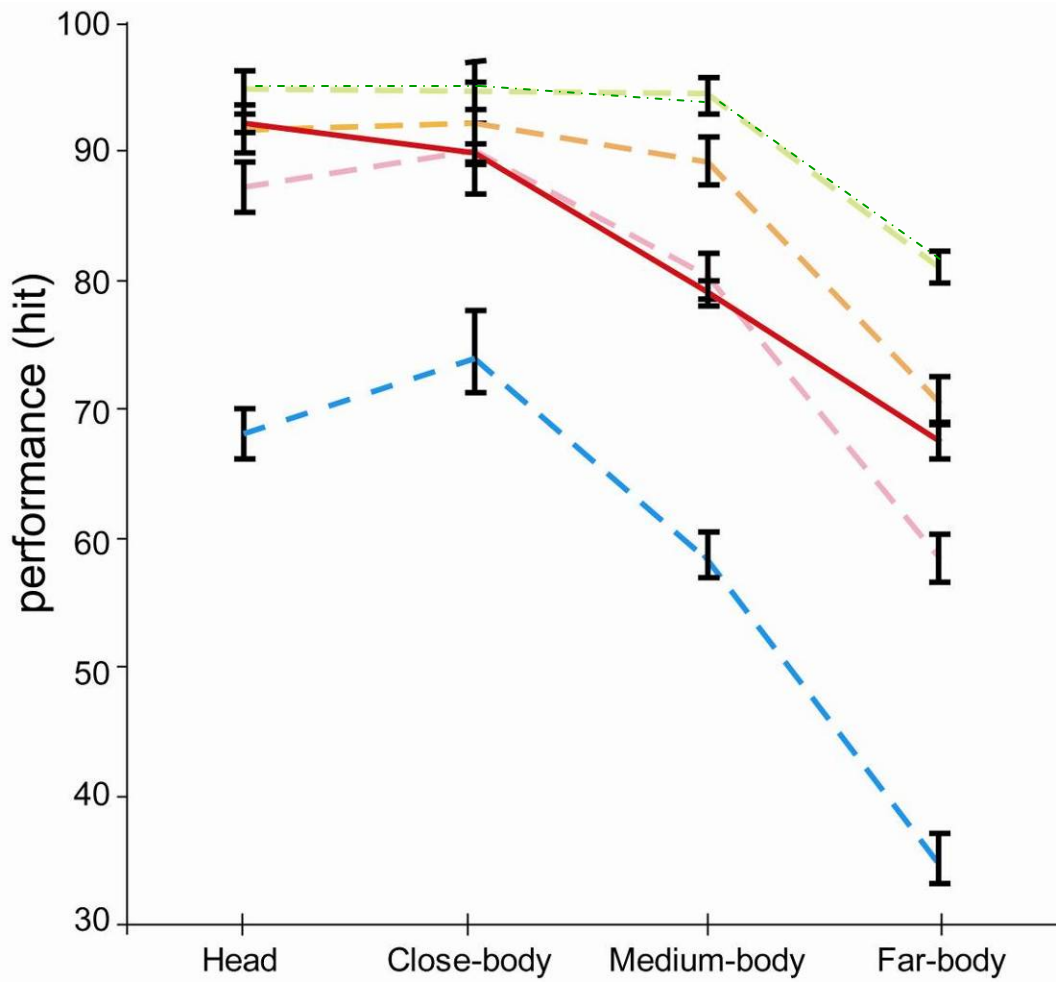


Serre Oliva Poggio 2007

What is next:
beyond feedforward models:
limitations

- Recognition in clutter is increasingly difficult
- Need for attentional bottleneck (Wolfe, 1994) perhaps in V4 (see Gallant and Desimone and models by Walther + Serre)
- Notice: this is a “novel” justification for the need of attention!

Limitations: beyond 50 ms: model not good enough



no mask condition

80 ms SOA (ISI=60 ms)

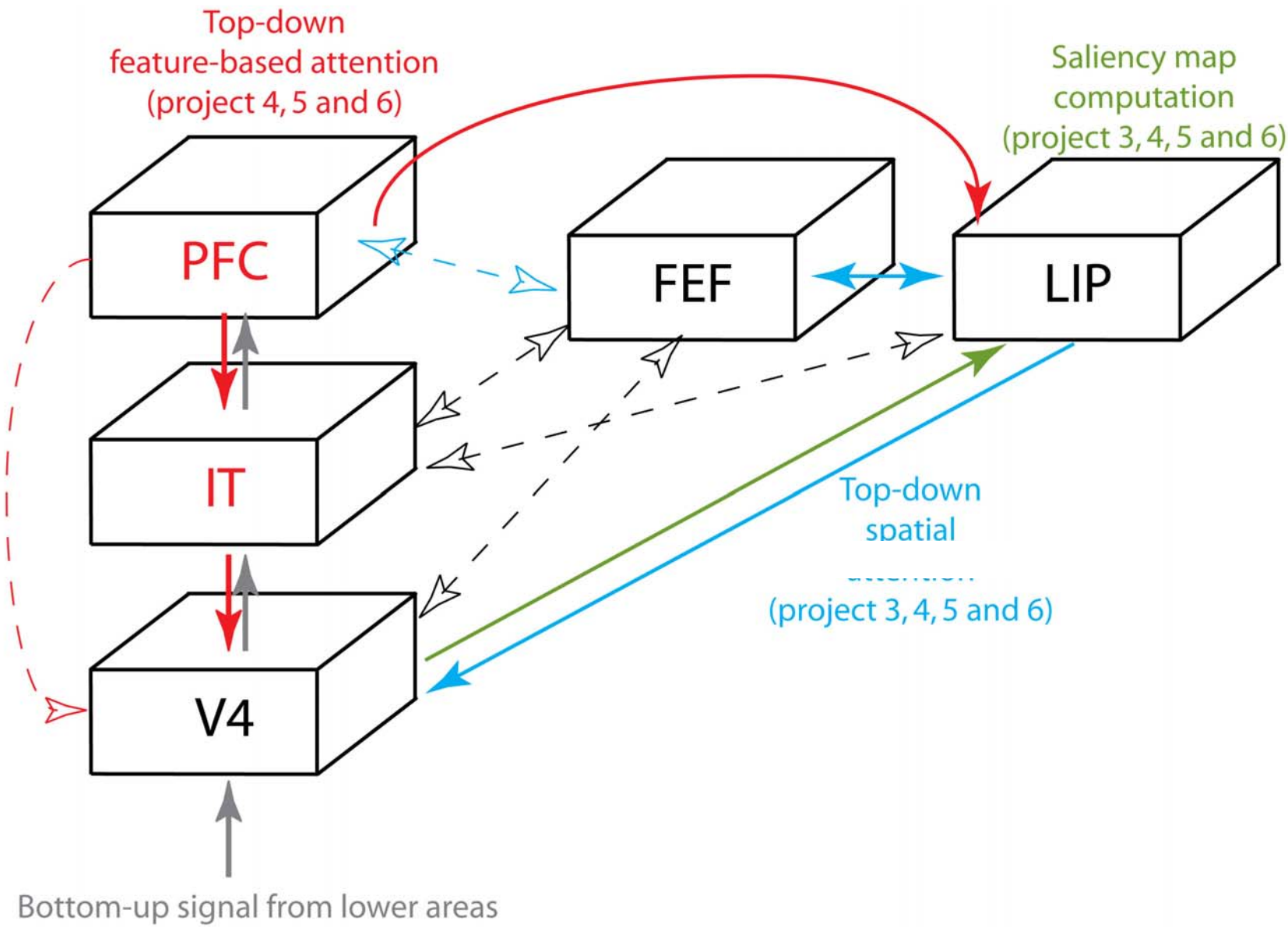
model

50 ms SOA (ISI=30 ms)

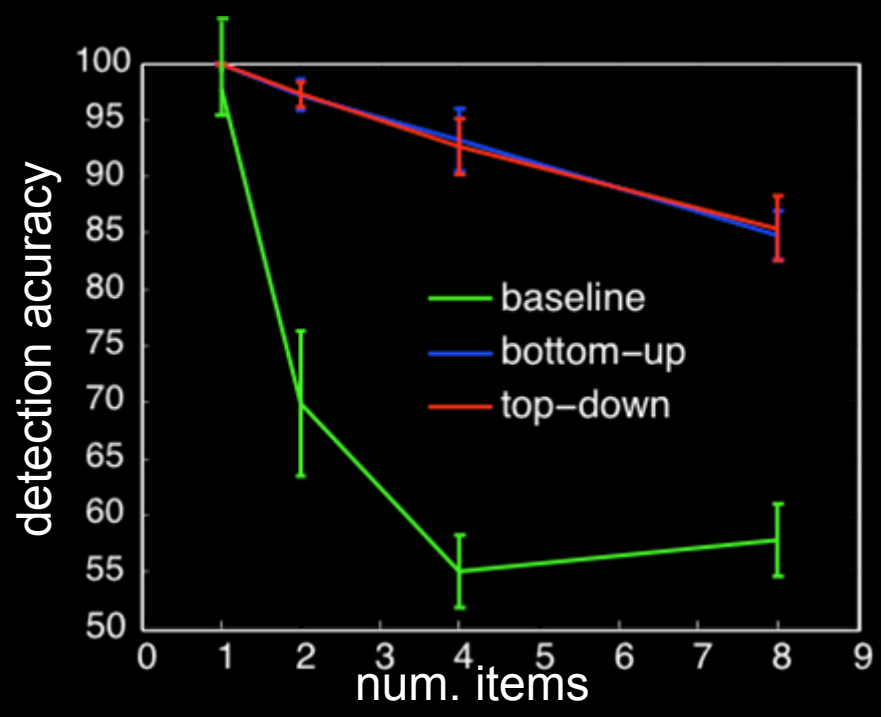
20 ms SOA (ISI=0 ms)

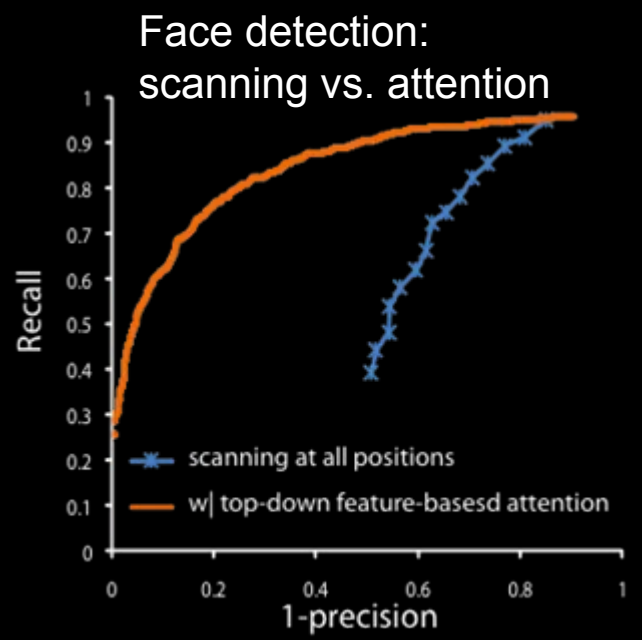
(Serre, Oliva and Poggio, PNAS, 2007)

Ongoing work….



Attention and cortical feedbacks

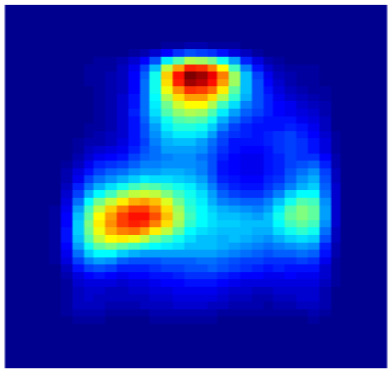
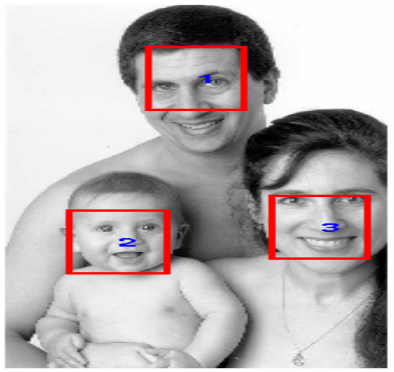


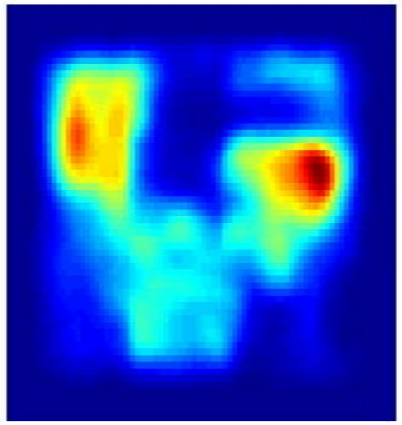


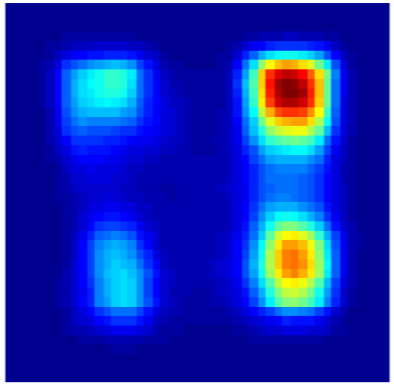
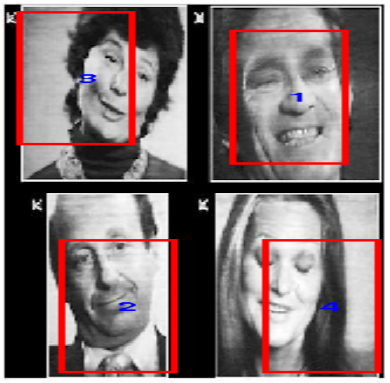
- Model implementation of Wolfe's guided search (1994)
- Parallel (feature-based top-down attention) and serial (spatial attention) to suppress clutter (Tsostos et al)

Chikkerur Serre Walther Koch & Poggio in prep

Example Results



What is next

- Image inference: attentional or Bayesian models?
- Why hierarchies? Beyond a model towards a theory
- Against hierarchies and the ventral stream: subcortical pathways

What is next: image inference, backprojections and attentional mechanisms

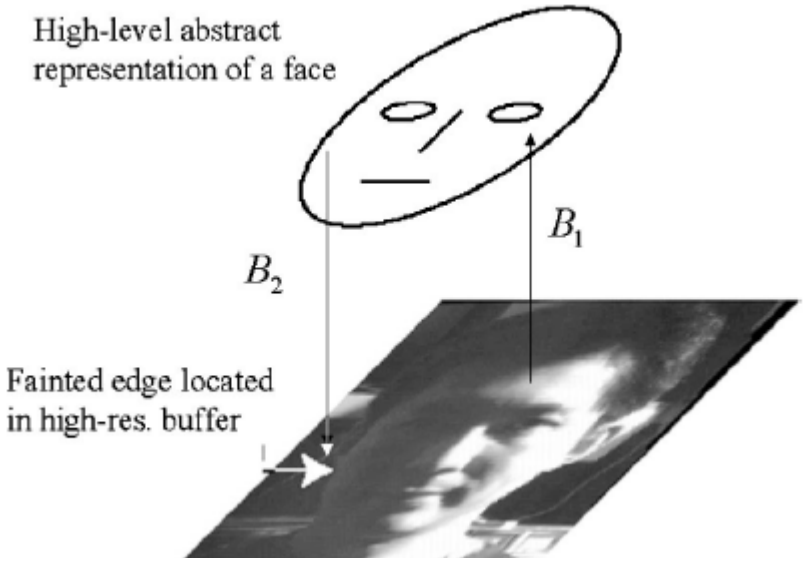
- Normal recognition by humans (for long times) is much better
- Normal vision is much more than categorization or identification: image understanding/inference/parsing

Attention-based models with high-level specialized routines

- Feedforward model + backprojections implementing featural and the spatial attention may improve recognition performance
- Backprojections also access/route information in/from lower areas to specific task-dependent routines in PFC (?). Open questions:
  - -- Which biophysical mechanisms for routing/gating?
  - -- Nature of routines in higher areas (eg PFC)?

Bayesian models

Analysis-by-synthesis models, eg probabilistic inference in the ventral stream: neurons represent conditional probabilities of the bottom-up sensory inputs given the top-down hypothesis and converge to globally consistent values



Lee and Mumford, 2003; Dean,2005 ;Rao, 2004; Hawkins, 2004; Ullman, 2007, Hinton, 2005

What is next

- Image inference: attentional or Bayesian models?
- Why hierarchies? Beyond a model, towards a theory
- Against hierarchies and the ventral stream: subcortical pathways (Bar et al., 2006, …)

Notices of the American Mathematical Society (AMS), Vol. 50, No. 5, 537-544, 2003.

The Mathematics of Learning: Dealing with Data

Tomaso Poggio and Steve Smale

How then do the learning machines described in the theory compare with brains?

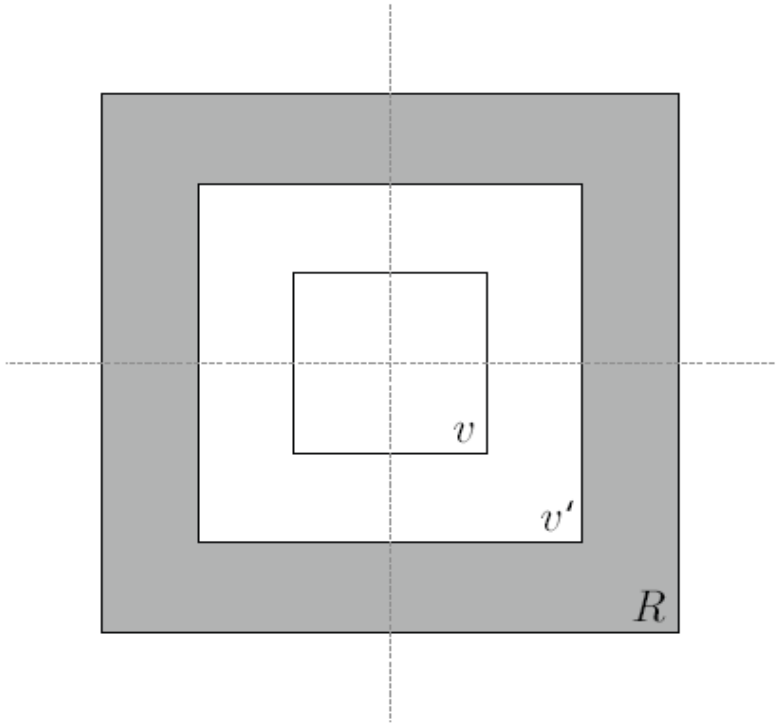
- One of the most obvious differences is the ability of people and animals to learn from very few examples.

- A comparison with real brains offers another, related, challenge to learning theory. The "learning algorithms" we have described in this paper correspond to one-layer architectures. Are hierarchical architectures with more layers justifiable in terms of learning theory?

- Why hierarchies? For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks.

- There may also be the more fundamental issue of *sample complexity*. Thus our ability of learning from just a few examples, and its limitations, may be related to the hierarchical architecture of cortex.

Formalizing the hierarchy: towards a theory



Axiom: $f \circ h : v \to [0,1]$ is in $Im(v)$ if $f \in Im(v')$ and $h \in H$,
that is *the restriction of an image is an image* and similarly for H'. Thus

$f \circ h : v \to [0,1] \in Im(v)$ if $f \in Im(v')$ and $h \in H$,
$f \circ h' : v' \to [0,1] \in Im(v')$ if $f \in Im(R)$ and $h' \in H'$.

We formulate the model in the following stages:

1. The process starts with some initial distance on $Im(v)$ provided by

$$d'_0(f,g) = d(f,g) = ||f-g||_p, \tag{1}$$

where $||\cdot||_p$ is an appropriate L_p norm $\left(||f||_p = \left(\int_v |f(x)|^p d\mu(x)\right)^{1/p}\right)$, for the space of functions $Im(v)$.

Then we define a first stage *Neural Similarity* as

$$N^1_t(f) = \min_{h\in H} d'_0(f\circ h, t), f \in Im(v') \tag{2}$$

Thus $N^1 : Im(v') \to \mathbf{R}^T_+$ can be defined[1] by $N^1(f)(t) = N^1_t(f)$.

We define the derived distance (with $||N^1(f)||_p = \left(\int_T |N^1_t|^p d\rho(t)\right)^{1/p}$) on $Im(v')$ as

$$d'_1(f,g) = ||N^1(f) - N^1(g)||_p \tag{3}$$

Since $N^1(f)$ and $N^1(g)$ are elements in $\mathbf{R}^T_+$, this norm makes sense (we use no L_p norm on $Im(v')$).

2. We now repeat the process by defining the second stage *Neural Similarity* as

$$N^2_{t'}(f) = \min_{h'\in H'} d'_1(f\circ h', t'), f \in Im(R), t' \in T'. \tag{4}$$

The new derived distance is now on $Im(R)$

$$d'_2(f,g) = ||N^2(f) - N^2(g)||_p. \tag{5}$$

Clearly this process could continue if appropriate higher level patches were defined.

Smale, S., T. Poggio, A. Caponnetto, and J. Bouvrie. Derived Distance: towards a mathematical theory of visual cortex, *CBCL Paper*, Massachusetts Institute of Technology, Cambridge, MA, November, 2007.

From a model to a theory: math results on unsupervised learning of invariances and of a dictionary of shapes from image sequences

The time evolution of the stimulus is modelled by a discrete time stationary process taking values in Im(Sq)

$$\mathbf{F} = \{F_\tau\}_{\tau \in \mathbb{Z}}$$

in the following $\mathbf{F}' = \{F'_\tau\}_{\tau \in \mathbb{Z}}$ is i.i.d. with $\mathbf{F}$.

Let us associate to every $v \in V(j)$ a denumerable partition of Im(Sq)

$$\mathcal{C}_v = \{C_v(k)\}_k$$

...

Proposition 2. *For every $v \in V(j)$, defining*

$$\sigma^* = 2 \left(1 + \sqrt{\frac{\mathbb{E}\left[K_v(R_v(F_0), R_v(F'_0))\right]}{\mathbb{E}\left[\|K_v(R_v(F_0)) - K_v(R_v(F_{\tau_j}))\|^2_{K_v}\right]}} \right)^{-1}$$

it holds

$$\mathrm{Err}_v(\sigma^*) \leq \check{\mathrm{E}}\mathrm{rr}_v.$$

Caponetto, Smale and Poggio, in preparation

(Obvious) caution remark!!!

There is still much to do before we understand vision…

and the brain!

Collaborators

T. Serre

- ❑ Model
  - ✓ A. Oliva
  - ✓ C. Cadieu
  - ✓ U. Knoblich
  - ✓ M. Kouh
  - ✓ G. Kreiman
  - ✓ M. Riesenhuber
- ❑ Comparison w| humans
  - ✓ A. Oliva
- ❑ Action recognition
  - ✓ H. Jhuang
- ❑ Attention
  - ✓ S. Chikkerur
  - ✓ C. Koch
  - ✓ D. Walther
- ❑ Computer vision
  - • S. Bileschi
  - • L. Wolf
- ❑ Learning invariances
  - • T. Masquelier
  - • S. Thorpe