Old HMAX
and face recognition

Tomaso Poggio
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
5. Beyond hierarchical models
WARNING: 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 a real theory
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
5. Beyond hierarchical models
Learning and Recognition in Visual Cortex:
what is where

Unconstrained visual recognition was a difficult problem (e.g., “is there an animal in the image?”)
Vision
A Computational Investigation into the Human Representation and Processing of Visual Information

David Marr
Foreword by Shimon Ullman
Afterword by Tomaso Poggio

David Marr's posthumously published *Vision* (1982) influenced a generation of brain and cognitive scientists, inspiring many to enter the field. In *Vision*, Marr describes a general framework for understanding visual perception and touches on broader questions about how the brain and its functions can be studied and understood. Researchers from a range of brain and cognitive sciences have long valued Marr's creativity, intellectual power, and ability to integrate insights and data from neuroscience, psychology, and computation. This MIT Press edition makes Marr's influential work available to a new generation of students and scientists.

In Marr's framework, the process of vision constructs a set of representations, starting from a description of the input image and culminating with a description of three-dimensional objects in the surrounding environment. A central theme, and one that has had far-reaching influence in both neuroscience and cognitive science, is the notion of different levels of analysis—in Marr's framework, the computational level, the algorithmic level, and the hardware implementation level.

Now, thirty years later, the main problems that occupied Marr remain fundamental open problems in the study of perception. *Vision* provides inspiration for the continu...
Vision: what is where
Vision: what is where

dorsal stream: “where”

ventral stream: “what”
The ventral stream...

Feedforward connections only?
...``solves” the problem

(if the mask forces feedforward processing)...

- $d'$~ standardized error rate
- the higher the $d'$, the better the performance

Model 82%

Human 80%

Serre Oliva & Poggio 2007
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
5. Beyond hierarchical models
Some personal history:
First step in developing a model:
learning to recognize 3D objects in IT cortex

Examples of Visual Stimuli

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!)

Regularization Network (GRBF) with Gaussian kernels

Prediction: neurons become view-tuned through learning

View Angle

Poggio & Edelman 1990
After human psychophysics (Buelthoff, Edelman, Tarr, Sinha, *to be added next year...*), which supports models based on view-tuned units...

... physiology!
Recording Sites in Anterior IT

...neurons tuned to faces are intermingled nearby....
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)
View-tuned cells: scale invariance (one training view only) motivates present model

Logothetis Pauls & Poggio 1995
• Gaussian centers (Gaussian Kernels) tuned to complex multidimensional features as composition of lower dimensional Gaussian

• What about tolerance to position and scale?
• Answer: hierarchy of invariance and tuning operations
Answer: the "HMAX" model

Riesenhuber & Poggio 1999, 2000
From HMAX to the present model

How the new version of the model evolved from the original one

1. **The two key operations:** Operations for selectivity and invariance, originally computed in a simplified and idealized form (i.e., a multivariate Gaussian and an exact max, see Section 2) have been replaced by more plausible operations, normalized dot-product and softmax.

2. **S1 and C1 layers:** In [Serre and Riesenhuber, 2004] we found that the S1 and C1 units in the original model were too broadly tuned to orientation and spatial frequency and revised these units accordingly. In particular at the S1 level, we replaced Gaussian derivatives with Gabor filters to better fit parafoveal simple cells’ tuning properties. We also modified both S1 and C1 receptive field sizes.

3. **S2 layers:** They are now learned from natural images. S2 units are more complex than the old ones (simple 2°—2 combinations of orientations). The introduction of learning, we believe, has been the key factor for the model to achieve a high-level of performance on natural images, see [Serre et al., 2002].

4. **C2 layers:** Their receptive field sizes, as well as range of invariances to scale and position have been decreased so that C2 units now better fit V4 data.

5. **S3 and C3 layers:** They were recently added and constitute the top-most layers of the model along with the S2b and C2b units (see Section 2 and above). The tuning of the S3 units is also learned from natural images.

6. **S2b and C2b layers:** We added those two layers to account for the bypass route (that projects directly from V1/V2 to PIT, thus bypassing V4 [see Nakamura et al., 1993]).
<table>
<thead>
<tr>
<th>Simple Cells</th>
<th>Model</th>
<th>Receptive Field Sizes</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$0.2^\circ - 1.1^\circ$</td>
<td>$= 0.1^\circ - 1.0^\circ$</td>
<td>[Schiller et al., 1976e; Hubel and Wiesel, 1965]</td>
</tr>
<tr>
<td>Complex Cells</td>
<td>$0.4^\circ - 1.6^\circ$</td>
<td>$= 0.2^\circ - 2.0^\circ$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Cells</th>
<th>Model</th>
<th>Peak Frequencies (cycles/deg)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cortex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>range: $1.6 - 9.8$</td>
<td>mean/med: $3.7/2.8$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean: $2.2$</td>
<td>range: $0.5 - 8.0$</td>
<td>[DeValois et al., 1982a])</td>
</tr>
<tr>
<td>Complex Cells</td>
<td>range: $1.8 - 7.8$</td>
<td>mean/med: $3.9/3.2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean: $3.2$</td>
<td>range: $0.5 - 8.0$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Cells</th>
<th>Model</th>
<th>Frequency Bandwidth at 50% Amplitude (cycles/deg)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cortex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>range: $1.1 - 1.8$</td>
<td>med: $1.45$</td>
<td>[DeValois et al., 1982a]</td>
</tr>
<tr>
<td>Complex Cells</td>
<td>range: $1.5 - 2.0$</td>
<td>med: $1.6$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bulk: $1.0 - 1.5$</td>
<td>range: $0.4 - 2.6$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Cells</th>
<th>Model</th>
<th>Frequency Bandwidth at 71% Amplitude (index)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cortex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>range: $44 - 58$</td>
<td>med: $55$</td>
<td>[Schiller et al., 1976d]</td>
</tr>
<tr>
<td>Complex Cells</td>
<td>range: $40 - 50$</td>
<td>med: $48$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bulk: $40 - 70$</td>
<td>range: $40 - 60$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Cells</th>
<th>Model</th>
<th>Orientation Bandwidth at 50% Amplitude (octaves)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cortex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>range: $38^\circ - 49^\circ$</td>
<td>med: $44^\circ$</td>
<td>[DeValois et al., 1982b]</td>
</tr>
<tr>
<td>Complex Cells</td>
<td>range: $27^\circ - 33^\circ$</td>
<td>med: $43^\circ$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bulk: $20^\circ - 90^\circ$</td>
<td>range: $20^\circ - 90^\circ$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Cells</th>
<th>Model</th>
<th>Orientation Bandwidth at 71% Amplitude (octaves)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cortex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>range: $27^\circ - 33^\circ$</td>
<td>med: $30^\circ$</td>
<td>[Schiller et al., 1976c]</td>
</tr>
<tr>
<td>Complex Cells</td>
<td>range: $27^\circ - 33^\circ$</td>
<td>med: $31^\circ$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bulk: $20^\circ - 90^\circ$</td>
<td>range: $20^\circ - 90^\circ$</td>
<td></td>
</tr>
</tbody>
</table>
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
5. Beyond hierarchical models
Vision: what is where

- Human Brain
  - $10^{10}$-$10^{11}$ neurons (~1 million flies)
  - $10^{14}$-$10^{15}$ synapses

- Neuron
  - Fundamental space dimensions:
    - fine dendrites: 0.1 μ diameter; lipid bilayer membrane: 5 nm thick; specific proteins: pump channels, receptors, enzymes
  - Fundamental time length: 1 msec

- Ventral stream in rhesus monkey
  - ~$10^9$ neurons in the ventral stream
    (350 $10^6$ in each hemisphere)
  - ~$15 \times 10^6$ neurons in AIT (Anterior InferoTemporal) cortex

Van Essen & Anderson, 1990
Neural Circuits

Source: Modified from Jody Culham’s web slides
Vision: what is where

Source: Lennie, Maunsell, Movshon
The ventral stream hierarchy: V1, V2, V4, IT
A gradual increase in the receptive field size, in the complexity of the preferred stimulus, in tolerance to position and scale changes

Kobatake & Tanaka, 1994
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
5. Beyond hierarchical models
The ventral stream hierarchy: V1, V2, V4, IT

A gradual increase in the receptive field size, in the complexity of the preferred stimulus, in tolerance to position and scale changes

Kobatake & Tanaka, 1994
Categorical judgments, decision making

120–160 ms

PFC

100–130 ms

MC

140–190 ms

PMC

30–50 ms

LGN

20–40 ms

Retina

80–100 ms

AIT

70–90 ms

High-level object descriptions, faces, objects

60–80 ms

V1

40–60 ms

V2

50–70 ms

V4

Simple visual forms, edges, corners

Motor command

Intermediate visual forms, feature groups, etc.

To spinal cord 160–220 ms

To finger muscle 180–260 ms

(Thorpe and Fabre-Thorpe, 2001)
V1: hierarchy of simple and complex cells

LGN-type cells  Simple cells  Complex cells

(Hubel & Wiesel 1959)
Recognition in the Ventral Stream: “classical model”

*Modified from (Gross, 1998)


[software available online with CNS (for GPUs)]

Recognition in Visual Cortex: “classical model”


- As a biological model of object recognition in the ventral stream – from V1 to PFC -- it is perhaps the most quantitatively faithful to known neuroscience data
## Two key computations, suggested by physiology

<table>
<thead>
<tr>
<th>Unit</th>
<th>Pooling</th>
<th>Computation</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td><img src="#" alt="Pooling" /></td>
<td>Selectivity / template matching</td>
<td>Gaussian-tuning / AND-like</td>
</tr>
<tr>
<td>Complex</td>
<td><img src="#" alt="Pooling" /></td>
<td>Invariance</td>
<td>Soft-max / or-like</td>
</tr>
</tbody>
</table>
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
Two operations (~OR, ~AND): disjunctions of conjunctions

- Tuning operation (Gaussian-like, AND-like)
  \[ y = e^{-|x-w|^2} \]
  or
  \[ y \sim \frac{x \cdot w}{|x|} \]

- Simple units

- Max-like operation (OR-like)
  \[ y = \max\{x_1, x_2, \ldots\} \]

- Complex units

Each operation ~microcircuits of ~100 neurons
Plausible biophysical implementations

- Max and Gaussian-like tuning can be approximated with same canonical circuit using shunting inhibition. Tuning (eg “center” of the Gaussian) corresponds to synaptic weights.
A plausible biophysical implementation for both Gaussian tuning (~AND) + max (~OR): normalization circuits with divisive inhibition (Kouh, Poggio, 2008; also RP, 1999; Heeger, Carandini, Simoncelli, …)

Recognition in Visual Cortex: circuits and biophysics

A canonical microcircuit of spiking neurons?
Simulation with spiking neurons and realistic synapses
Figure 3: Mean response of max circuit depicted in Fig. 2 over 50 runs for all possible combinations of 0, 50, 75 and 100 spikes per input packet, plotted against the desired (true) maximum of the inputs (left). Histogram of all outputs (spike count in output packet) for three cases (right). The true maximum of the inputs is 50, 75 and 100 spikes, respectively (top to bottom).

Figure 4: Output (spike count in output packet) of a one-dimensional Gaussian-like tuning circuit tuned to 50 a spike packet input (left). Output (spike count in output packet) of the two-dimensional tuning circuit depicted in Fig. 2 tuned to the combination of two 50 spike packet inputs (right).
Basic circuit is closely related to other models

<table>
<thead>
<tr>
<th>Operation</th>
<th>(Steady-State) Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canonical</strong></td>
<td>[ y = \frac{\sum_{i=1}^{n} w_i x_i^p}{k + \left( \sum_{i=1}^{n} x_i^q \right)^{\frac{1}{p}}} ] (1)</td>
</tr>
<tr>
<td><strong>Energy Model</strong></td>
<td>[ y = \sum_{i=1}^{2} x_i^2 ] (2)</td>
</tr>
<tr>
<td><strong>Gaussian-like</strong></td>
<td>[ y = \frac{\sum_{i=1}^{n} w_i x_i^3}{k + \sum_{i=1}^{n} x_i^2} ] (4)</td>
</tr>
<tr>
<td><strong>Max-like</strong></td>
<td>[ y = \frac{\sum_{i=1}^{n} x_i^3}{k + \sum_{i=1}^{n} x_i^2} ] (5)</td>
</tr>
</tbody>
</table>

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)

Adelson and Bergen (see also Hassenstein and Reichardt, 1956)

Of the same form as model of MT (Rust et al., Nature Neuroscience, 2007)
Task-specific circuits (from IT to PFC?)
- **Supervised** learning: ~ classifier

Overcomplete dictionary of “templates” ~ image “patches” ~ ~ “parts” is learned during an **unsupervised** learning stage (from ~10,000 natural images) by tuning S units.

see also (Foldiak 1991; Perrett et al 1984; Wallis & Rolls, 1997; Lewicki and Olshausen, 1999; Einhauser et al 2002; Wiskott & Sejnowski 2002; Spratling 2005)
Recognition in Visual Cortex: learning
(from Serre, 2007)
Start with S2 layer

- Units are organized in n feature maps
- Database ~1,000 natural images

At each iteration:
- Present one image
- Learn k feature maps
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.

Image “moves” (looming and shifting)

Weight vector \( w \) is copied to all units in feature map 1 (across positions and scales)
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)
Recognition in Visual Cortex: learning

Sample S2 Units Learned *(from Serre, 2007)*
Neurons in monkey visual area V2 encode combinations of orientations
Akiyuki Anzai, Xinmiao Peng & David C Van Essen
Comparison w/ V4

Tuning for curvature and boundary conformations?
**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 different positions and scales
- A prediction to be tested: S2 units in V2 and C2 units in V4?
A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4

Jay Hegdé and David C. Van Essen
Beyond C2 units

- Units increasingly complex and invariant
- S3/C3 units:
  - Combination of V4-like units with different selectivities
  - Dictionary of ~1,000 features = num. columns in IT (Fujita 1992)
A loose hierarchy

• Bypass routes along with main routes:
  • From V4 to TE (bypassing TEO) (Desimone et al 1980; Saleem et al 1992)
• “Replication” of simpler selectivities from lower to higher areas
• Rich dictionary of features – across areas -- with various levels of selectivity and invariance
Readings on the work with many relevant references

A detailed description of much of the work is in the “supermemo” at

Other recent publications and references can be found at
http://cbcl.mit.edu/publications/index-pubs.html
The most recent version of this straightforward class of models is consistent with many data at different levels -- from the **computational to the biophysical level**.

Being testable across all these levels is a high bar and an important one (too easy to develop models that explain one phenomenon or one area or one illusion...these models overfit the data, they are not scientific)
Recognition in Visual Cortex: model accounts for psychophysics
Hierarchical feedforward models of the ventral stream

Rapid Categorization: mask should force visual cortex to operate in feedforward mode

Thorpe et al 1996; Van Rullen & Koch 2003; Bacon-Mace et al 2005
Hierarchical feedforward models of the ventral stream
Recognition in Visual Cortex: model accounts for psychophysics

Feedforward Models: “predict” rapid categorization (82% model vs. 80% humans)

Image-by-image correlation: around 73% for model vs. humans
Hierarchical model of recognition in visual cortex

- Image-by-image correlation:
  - Heads: $\rho=0.71$
  - Close-body: $\rho=0.84$
  - Medium-body: $\rho=0.71$
  - Far-body: $\rho=0.60$
Agreement of model with IT Readout data
Reading-out the neural code in AIT

77 objects, 8 classes

Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, Science, Nov 4, 2005
Recording at each recording site during passive viewing

- 77 visual objects
- 10 presentation repetitions per object
- Presentation order randomized and counter-balanced

Time sequence: 100 ms 100 ms
Agreement of model w/ IT Readout data
Training a classifier on neuronal activity.

From a set of data (vectors of activity of n neurons (x) and object label (y))

\[
\{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\}
\]

Find (by training) a classifier eg a function f such that

\[f(x) = \hat{y}\]

is a good predictor of object label y for a future neuronal activity x
Decoding the Neural Code ... population response (using a classifier)

Learning from \((x, y)\) pairs

\[ y \in \{1, \ldots, 8\} \]
From neuronal population activity…

…a classifier can decode and guess what the monkey was seeing…

Categorization
- Toy
- Body
- Human Face
- Monkey Face
- Vehicle
- Food
- Box
- Cat/Dog

Video speed: 1 frame/sec
Actual presentation rate: 5 objects/sec
80% accuracy in read-out from ~200 neurons

So...experimentally we can decode the brain’s code and read-out from neural activity what the monkey is seeing.

We can also read-out with similar results from the model !!!
A result (C. Hung, et al., 2005): very rapid read-out of object information rapid (80-100 ms from onset of stimulus)

Information represented by population of neurons over very short times (over 12.5ms bin)

Very strong constraint on neural code (not firing rate). Consistent with our IF circuits for max and tuning
It turns out that the model agrees with IT data: we can decode from model units as well as from IT.
A result (C. Hung, et al., 2005): very rapid read-out of object information rapid (80-100 ms from onset of stimulus).

Information represented by population of neurons over very short times (over 12.5 ms bin).

Very strong constraint on neural code (not firing rate). Consistent with our IF circuits for max and tuning.
Agreement of model w/ IT Readout data
Reading out category and identity invariant to position and scale

Hung Kreiman Poggio DiCarlo 2005

Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005
Agreement of Model w\| IT Readout data

Reading out category and identity “invariant” to position and scale

Hung, et al. 2005; Serre et al., 2005
• 70/30 train/test (20 splits)
• 64 randomly selected C3/C2b features
  – to match 64 recording sites
• **Scale:** 77.2 ± 1.25% vs. ~63% (physiology)
• **Location:** 64.9 ± 1.44% vs. ~65% (physiology)
• **Categorization:** 71.6 ± 0.91% vs. ~77% (physiology)
Models of the \textit{ventral stream} in cortex perform well compared to engineered computer vision systems (in 2006) on several databases.

Bileschi, Wolf, Serre, Poggio, 2007
Model extension to the dorsal stream: Recognition of actions

Thomas Serre, Hueihan Jhuang & Tomaso Poggio collaboration with David Sheinberg at Brown University
Behavioral analyses of mouse behavior needed to:
- Assess functional roles of genes
- Validate models of mental diseases
- Help assess efficacy of drugs

Automated quant system to help:
- Limit subjectivity of human intervention
- 24/7 home-cage analysis of behavior
- 24/7 monitoring of animal well-being
Models of the *dorsal stream* in cortex lead to better systems for action recognition in videos: automatic phenotyping of mice.

Hierarchical model of recognition: action recognition, ventral + dorsal stream (Giese and Poggio 2003);
Models of cortex lead to better systems for action recognition in videos: automatic phenotyping of mice

<table>
<thead>
<tr>
<th>Performance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human agreement</td>
<td>72%</td>
</tr>
<tr>
<td>proposed system</td>
<td>77%</td>
</tr>
<tr>
<td>commercial system</td>
<td>61%</td>
</tr>
<tr>
<td>chance</td>
<td>12%</td>
</tr>
</tbody>
</table>

Performance of the proposed system was compared to human agreement and commercial systems. The proposed system achieved 77%, significantly higher than the 72% agreement with humans and 61% for the commercial system, with a chance level of 12%.

*Jhuang, Garrote, Yu, Khilnani, Poggio, Mutch Steele, Serre, Nature Communications, 2010*
Model “works”: it accounts for physiology

Hierarchical Feedforward Models: is consistent with or predict neural data

V1:
Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
MAX-like operation in subset of complex cells (Lampl et al 2004)

V2:
Subunits and their tuning (Anzai, Peng, Van Essen 2007)

V4:
Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)
MAX-like operation (Gawne et al 2002)
Two-spot interaction (Freiwald et al 2005)
Tuning for boundary conformation (Pasupathy & Connor 2001, Cadieu, Kouh, Connor et al., 2007)
Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)

IT:
Tuning and invariance properties (Logothetis et al 1995, paperclip objects)
Read out results (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)
Motivation: theory is needed!

Hierarchical, Hubel and Wiesel (HMAX-type) models work well, as model of cortex and as computer vision systems but...why? and how can we improve them?

Similar convolutional networks called deep learning networks (LeCun, Hinton,...) are unreasonably successful in vision and speech (ImageNet+Timit)... why?
Recognition in Visual Cortex: computation and mathematical theory

For 10 years+...

I did not manage to understand how model works....

we need theories -- not only models!
What do hierarchical architectures compute? How? How do they develop?

THE COMPUTATIONAL MAGIC OF THE VENTRAL STREAM: TOWARDS A THEORY

...answer in a few weeks!

Tomaso Poggio*† (section 4 with Jim Mutch*; appendix 7.2 with Joel Leibo* and appendix 7.9 with Lorenzo Rosasco†)
* CBCL, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA
† Istituto Italiano di Tecnologia, Genova, Italy
Efficient software implementation: a GPU-based framework for simulating cortically-organized networks

(\textbf{CNS: available on our Web site})

Feedforward object recognition (static CBCL model):
- 256x256 input, 12 orientations, 4,075 “S2” features.
- Best CPU-based implementation: 28.2 sec/image.
- CNS (on NVIDIA GTX 295): 0.291 sec/image \((97\times\text{speedup})\).

Action recognition in streaming video:
- 8 9x9x9 spatiotemporal filters, 300 S2 features.
- Best CPU-based implementation: 0.55 fps.
- CNS: 32 fps \((58\times\text{speedup})\).

Spiking neuron simulation (dynamic model):
- 9,808 Hodgkin-Huxley neurons and 330,295 synapses.
- 310,000 simulated time steps required 57 seconds.
Extension to attention: dealing with clutter

see also Broadbent 1952 1954; Treisman 1960; Treisman & Gelade 1980; Duncan & Desimone 1995; Wolfe, 1997; Tsotsos and many others
Readings on the work with many relevant references

A detailed description of much of the work is in the “supermemo” at

Other recent publications and references can be found at
http://cbcl.mit.edu/publications/index-pubs.html
A view of FaceProcessing from the vantage point of HMAX

(with Cheston Tan)
Part I: Understanding Holism
Who is shown in the top half?
Versions of CFE task

Naming

Discrimination
Hypothesis: holistic processing is due to large templates

- Normal template matching
- Same underlying mechanisms
- No (less) holism for small templates
HMAX model

Riesenhuber & Poggio (1999); Mutch & Lowe (2008)
Activity of “V1 complex cells”
HMAX model

S1
Local max
4 orientations
Image

C1
Local max

Riesenhuber & Poggio (1999); Mutch & Lowe (2008)
S2 template = 4 x n x n

S2_{fine} templates (finer scale)
S2 template = 4 x n x n

S2_{coarse} templates (coarser scale)
S2 template = 4 x n x n

Large, coarse vs. small, fine

In reality, probably a mixture and/or continuum
HMAX model

Riesenhuber & Poggio (1999); Mutch & Lowe (2008)
Two “processing styles”

Riesenhuber & Poggio (1999); Mutch & Lowe (2008)
Hypothesis: holistic processing is due to large templates

- Normal template matching
- Same underlying mechanisms
- No (less) holism for small templates
Holism at single-unit level?

Kobatake & Tanaka (1994)
Task for CFE

Ignore bottom halves
Compare top halves
Simulation details

• Not part of model per se...
  – make reasonable assumptions
  – not trying to solve attention and decision-making

• Attentional modulation
  – bottom pixels attenuated

• Distance metric
  – euclidean distance

• Threshold
  – if distance < threshold, then “same”
Simulation results

Large, coarse features
Part III: Discussion
Holism without wholes
Why large, coarse templates?

- Stimulus properties, genetics, ubiquity, task demands, social demands, infant visual acuity, etc …
- Stimulus properties: first-order configuration
- Genetics + infant visual acuity
- Ubiquity + infant visual acuity
- etc …
In-principle proofs:

1) Face and object processing could share the same underlying mechanisms

2) There may be just one type of face processing
Computational performance: example faces
Labeled Faces in the Wild

Contains 13,233 images of 5,749 people
Accuracy (%)

- LFW - no outside data used & no alignment

Our model: 87.6%
- APEM: 81.7%
- Li et al. (2013): 79.08%
- Sanderson et al. (2009): 72.95%
- Rahimzadeh et al. (2013): 79.08%
Class specific pose invariance for faces
Testing computational performance

Labeled Faces in the Wild: contains 13,233 images of 5,749 people

Face verification: same-di fferent matching
Computational performance: example faces

Pubfig

- Originally, 58,797 images of 200 people
- Only ~21000 left now
Computational performance: example faces
Labeled Faces in the Wild

Contains 13,233 images of 5,749 people
Testing computational performance

(A) PIPELINE
1. Detection
2. Alignment
3. Recognition

(B) PERFORMANCE

<table>
<thead>
<tr>
<th></th>
<th>PubFig83</th>
<th>MIT-450</th>
<th>LFW</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>63.8</td>
<td>65.3</td>
<td>73.3</td>
</tr>
<tr>
<td>HOG Signatures</td>
<td>73.7</td>
<td>74.7</td>
<td>83.5</td>
</tr>
</tbody>
</table>

(C) ROC CURVES

Face verification: same-different matching

J. Leibo, Q. Liao
Performance Summary Face verification

- Pubfig State-of-the-art: 78.65% (original training and testing set)
- Our current performance:
  HOG based: ~78.3%
  LBP based: ~78.5%
  LBP + HOG based: ~80.5%
- We did not touch their training data at all.

J. Leibo, Q. Liao