The computational magic of the ventral stream

tomaso poggio
McGovern Institute, I2, CBCL, BCS, CSAIL
MIT
(with Jim Mutch, Joel Leibo, Lorenzo Rosasco)

A dream:
I have a theory of what the ventral stream does and how; it can explain why evolution chose a hierarchical architecture and how computations determine properties of cells specific for each visual area.

Very preliminary, provocative, probably completely wrong theory, but if true... what else do we want from computational neuroscience?
LEARNING THEORY
+
ALGORITHMS

ENGINEERING
APPLICATIONS

COMPUTATIONAL
NEUROSCIENCE:
models+experiments


Vision: what is where

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
The ventral stream

Movshon et al.

Desimone & Ungerleider 1989

Monday, August 29, 2011
Recognition in the Ventral Stream: classical, “standard” model (of immediate recognition)  

*Modified from (Gross, 1998)

[software available online with CNS (for GPUs)]

Model “works”: it accounts for bits of physiology + psychophysics

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)
Model “works”: it performs well at computational level

Models of the *ventral stream* in cortex perform well compared to engineered computer vision systems (in 2006) on several databases

Bileschi, Wolf, Serre, Poggio, 2007
Model “works”: it performs well at computational level

Models of the ventral stream in cortex perform well compared to engineered computer vision systems (in 2006) on several databases

Bileschi, Wolf, Serre, Poggio, 2007
Model “works”: it performs well at computational level

Models of the *ventral stream* in cortex perform well compared to engineered computer vision systems (in 2006) on several databases.

Bileschi, Wolf, Serre, Poggio, 2007
Recognition in Visual Cortex: computation and mathematical theory

For 10 years+...

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

we need a theory -- not only a model!
Why do hierarchical architectures work?

If features do not matter....what does?
THE COMPUTATIONAL MAGIC OF THE VENTRAL STREAM: TOWARDS A THEORY

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
What is the main difficulty of object recognition? Geometric transformations or intraclass variability?

- We have “factored out” viewpoint, scale, translation: is it now still “difficult” for a classifier to categorize dogs vs horses?
Are transformations the main difficulty for biological object recognition?
6 main ideas, steps in the theory

1. The computational goal of the ventral stream is to learn transformations (affine in $R^2$) during development and be invariant to them (McCulloch+Pitts 1945, Hoffman 1966, Jack Gallant, 1993, ...)

2. A memory-based module, similar to simple-complex cells, can learn from a video of an image transforming (e.g., translating) to provide a signature to any single image of any object that is invariant under the transformation (*Invariance Lemma*).

3. Since the group $\text{Aff}(2,R)$ can be factorized as a semidirect product of translations and $\text{GL}(2, R)$, storage requirements of the memory-based module can be reduced by order of magnitudes (*Factorization Theorem*). I argue that this is the evolutionary reason for hierarchical architecture in the ventral stream.
6 main ideas, steps in the theory

4. Evolution selects translations by using small apertures in the first layer (the Stratification Theorem). Size of receptive fields in the sequence of ventral stream area automatically selects specific transformations (translations, scaling, shear and rotations).

5. If Hebb-like rule active at synapses then the storage requirements can be reduced further: tuning of cell in each layer -- during development -- will mimic the spectrum (SVD) of stored templates (Linking Theorem).

6. At the top layers invariances learned for “places” and for class-specific transformations such changes of expression of a face or rotation in depth of a face or different poses of a body ====> class-specific patches
Stratification

- Aperture size
- Objet class
- Transformations
  - Specific class
    - Specific complex
      - Notations
      - Expansions...
  - Generic
- Translations
SVD (of templatebook) for (x) translations from natural images
SVD (of templatebook) for \( y \) translations noise

Jim Mutch

Monday, August 29, 2011
These are predicted features orthogonal to trajectories of Lie group: for translation $(x, y)$, expansion, rotation

This is from Hoffman, 1966 cited by Jack Gallant
Implications

• Image prior, shape features, natural images priors are “irrelevant”

• The type of transformation that are learned from visual experience depend on the size (measured in terms of wavelength) and thus on the area (layer in the models) -- assuming that the aperture size increases with layers

• The mix of transformations learned determine the properties of the receptive fields -- oriented bars in V1+V2, radial and spiral patterns in V4 up to class specific tuning in AIT (eg face tuned cells)
Some predictions

• Invariance to small transformations in early areas (e.g., translations in V1) may underly stability of visual perception (suggested by Stu Geman)
• Each cell's tuning properties are shaped by visual experience of image transformations during developmental and adult plasticity
• Simple cells are likely to be the same population as complex cells, arising from different convergence of the Hebbian learning rule. The input to complex "complex" cells are dendritic branches with simple cell properties
• Class-specific transformations are learned and represented at the top of the ventral stream hierarchy; thus class-specific modules -- such as faces, places and possibly body areas -- should exist in IT
• The type of transformations that are learned from visual experience depend on the size of the receptive fields and thus on the area (layer in the models) -- assuming that the size increases with layers
• The mix of transformations learned in each area influences the tuning properties of the cells -- oriented bars in V1+V2, radial and spiral patterns in V4 up to class specific tuning in AIT (e.g., face tuned cells).
Collaborators in recent work

J. Leibo, J. Mutch, L. Isik,  
S. Ullman, S. Smale, L. Rosasco,  
H. Jhuang, C. Tan, N. Edelman, E. Meyers, B. Desimone  
O Lewis, S Voinea