Object Recognition

Mark van Rossum

School of Informatics, University of Edinburgh

January 15, 2018

[0] Based on slides by Chris Williams. Version: January 15, 2018

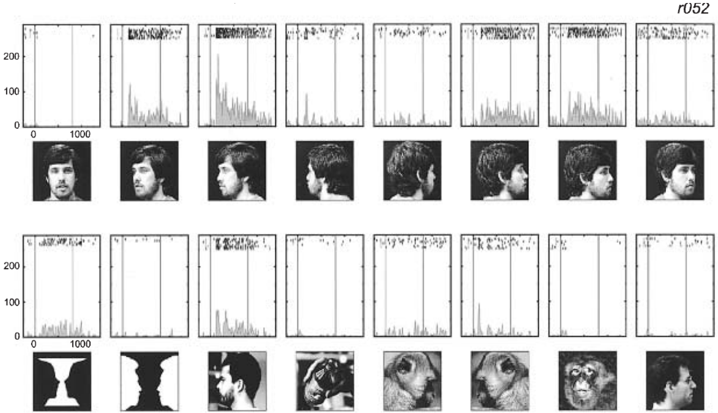
Overview

- Neurobiology of Vision
- Computational Object Recognition: What’s the Problem?
- Fukushima’s Neocognitron
- HMAX model and recent versions
- Other approaches

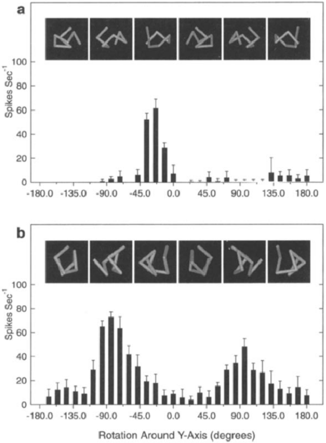
Neurobiology of Vision

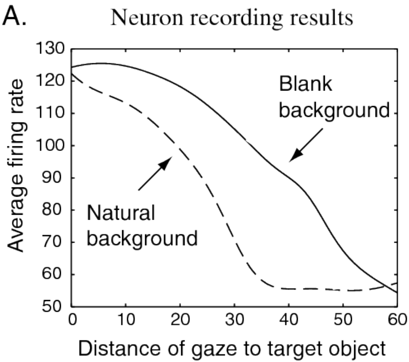
- WHAT pathway: V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ IT
- WHERE pathway: V1 $\rightarrow$ V2 $\rightarrow$ V3 $\rightarrow$ MT/V5 $\rightarrow$ parietal lobe
- IT (Inferotemporal cortex) has cells that are
  - Highly selective to particular objects (e.g. face cells)
  - Relatively invariant to size and position of objects, but typically variable wrt 3D view
- What and where information must be combined somewhere

Invariances in higher visual cortex



[?]





Left: partial rotation invariance [**?**].
Right: clutter reduces translation invariance [**?**].

thways/index.html

Computational Object Recognition

- The big problem is creating *invariance* to scaling, translation, rotation (both in-plane and out-of-plane), and partial occlusion, while at the same time being selective.
- What about a back-propagation network that learns some function $f(I_{x,y})$?
    - Large input dimension, need enormous training set
    - No invariances a priori
- Objects are not generally presented against a neutral background, but are embedded in *clutter*
- Tasks: object-*class* recognition, specific object recognition, localization, segmentation, ...

Some Computational Models

Two extremes:

- Extract 3D description of the world, and match it to stored 3D structural models (e.g. human as generalized cylinders)
- Large collection of 2D views (templates)

Some other methods

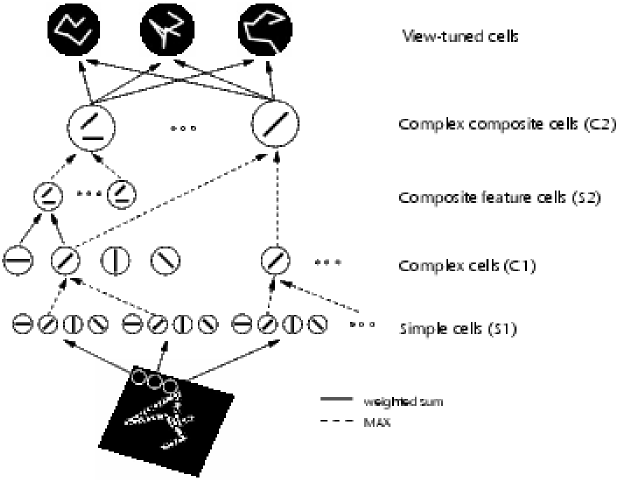
- 2D structural description (parts and spatial relationships)
- Match image features to model features, or do pose-space clustering (Hough transforms))
  - What are good types of features?
- Feedforward neural network
- Bag-of-features (no spatial structure; but what about the “binding problem”?)
- Scanning window methods to deal with translation/scale

Fukushima’s Neocognitron

[**?**, **?**]

- To implement location invariance, “clone” (or replicate) a detector over a region of space, and then pool the responses of the cloned units
- This strategy can then be repeated at higher levels, giving rise to greater invariance
- See also [**?**], *convolutional* neural networks

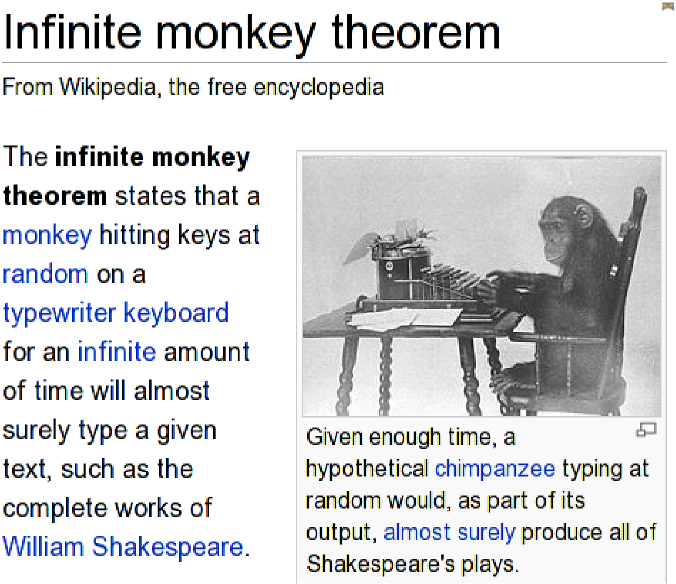
HMAX model


View-tuned cells
Complex composite cells (C2)
Composite feature cells (S2)
Complex cells (C1)
Simple cells (S1)
weighted sum
MAX

[?]

HMAX model

- S1 detectors based on Gabor filters at various scales, rotations and positions
- S-cells (simple cells) convolve with local filters
- C-cells (complex cells) pool S-responses with maximum
- No learning between layers
- Object recognition: Supervised learning on the output of C2 cells.



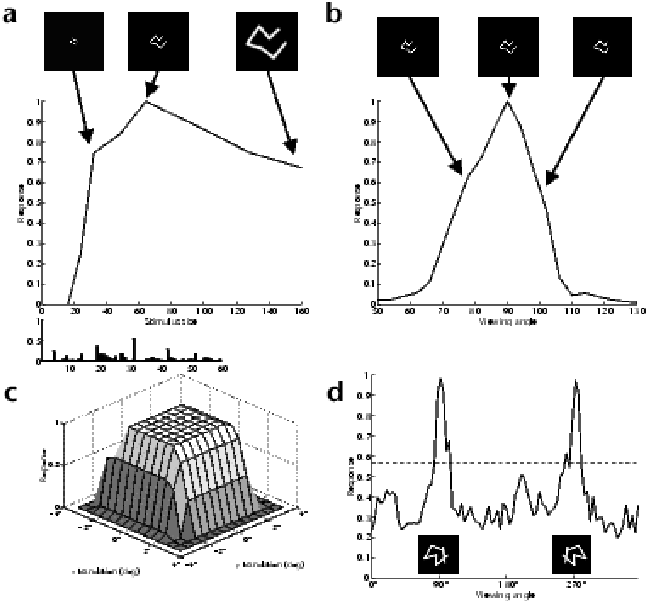
Infinite monkey theorem

From Wikipedia, the free encyclopedia

The **infinite monkey theorem** states that a monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type a given text, such as the complete works of William Shakespeare.

Given enough time, a hypothetical chimpanzee typing at random would, as part of its output, almost surely produce all of Shakespeare's plays.

Rather than learning, take refuge in having many, many cells. (Cover, 1965)*A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is*

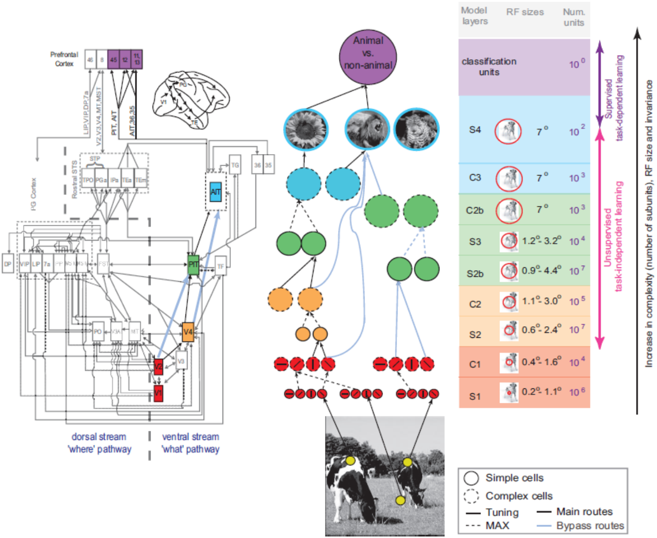

a
b
c
d
Response
Stimulus size
0
20
40
60
80
100
120
140
160
10
20
30
40
50
60
0.5
Viewing angle
50
60
70
80
90
100
110
120
130
x translation (deg)
y translation (deg)
0°
90°
180°
270°
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

[?]

HMAX model: Results

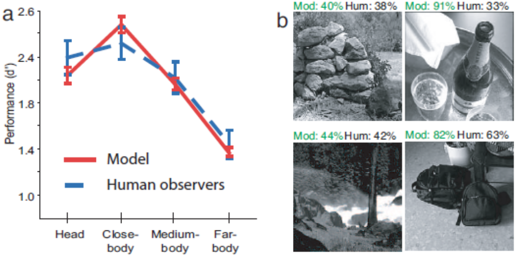
- “paper clip” stimuli
- Broad tuning curves wrt size, translation
- Scrambling of the input image does not give rise to object detections: not all conjunctions are preserved

More recent version



Model layers	RF sizes	Num. units
classification units		10^0
S4	7°	10^2
C3	7°	10^3
C2b	7°	10^3
S3	1.2°- 3.2°	10^4
S2b	0.9°- 4.4°	10^7
C2	1.1°- 3.0°	10^5
S2	0.6°- 2.4°	10^7
C1	0.4°- 1.6°	10^4
S1	0.2°- 1.1°	10^6

[?]



- Use real images as inputs
- S-cells convolution,e.g. $h = (\frac{\sum_i w_i x_i}{\kappa + \sqrt{\sum_i w_i^2}})$, $y = g(h)$.
- C-cell soft-max pooling $h = \frac{\sum x_i^{q+1}}{\kappa + \sum_k x_i^q}$
  (some support from biology for such pooling)
- Some unsupervised learning between layers [**?**]

Results

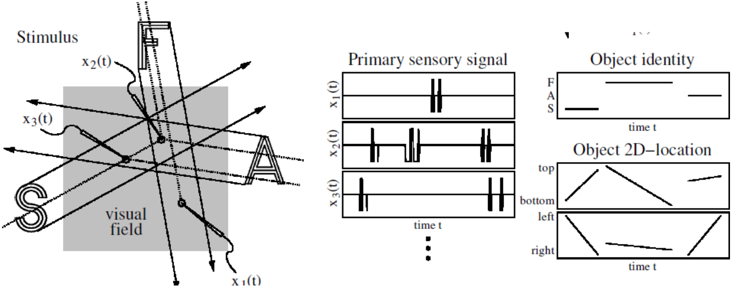
- Localization can be achieved by using a sliding-window method
- Claimed as a model on a “rapid categorization task”, where back-projections are inactive
- Performance similar to human performance on flashed (20ms) images
- The model doesn’t do segmentation (as opposed to bounding boxes)

Learning invariances

- Hard-code (convolutional network)
  http://yann.lecun.com/exdb/lenet/
- Supervised learning (show various sample and require same output)
- Use temporal continuity of the world. Learn invariance by seeing object change, e.g. it rotates, it changes colour, it changes shape.
  Algorithms: trace rule[**?**]
  E.g. replace
  $\Delta w = x(t).y(t)$ with $\Delta w = x(t).\tilde{y}(t)$
  where $\tilde{y}(t)$ is temporally filtered $y(t)$.
- Similar principles: VisNet [**?**], Slow feature analysis.

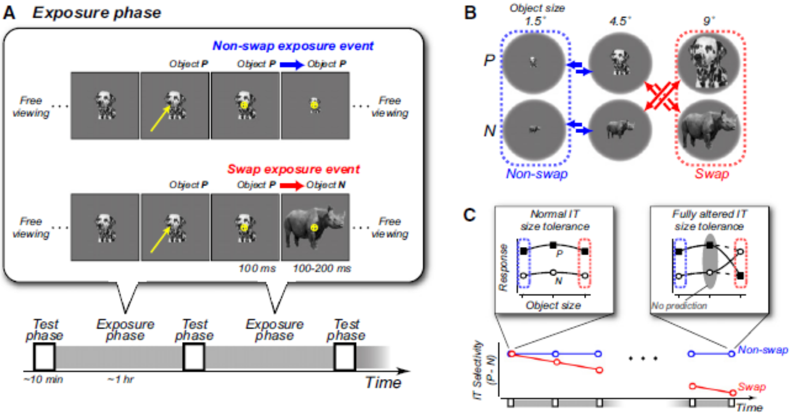
Slow feature analysis

Find slow varying features, these are likely relevant [**?**]



Find output y for which: $\langle (\frac{dy(t)}{dt})^2 \rangle$ minimal,
while $\langle y \rangle = 0, \langle y^2 \rangle = 1$

Experiments: Altered visual world [**?**]



A different flavour Object Recognition Model

[**?**]

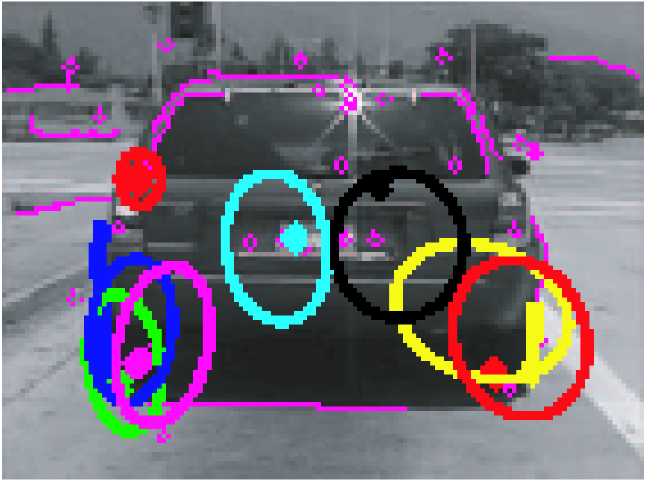
- Preprocess image to obtain interest points
- At each interest point extract a local image descriptor (e.g. Lowe's SIFT descriptor). These can be clustered to give discrete “visual words”
- $(w_i, \mathbf{x}_i)$ pair at each interest point, defining visual word and location
- Define a *generative* model. Object has instantiation parameters $\boldsymbol{\theta}$ (location, scale, rotation etc)
- Object also has *parts*, indexed by z

$$p(w_i, \mathbf{x}_i|\boldsymbol{\theta}) = \sum_{j=0}^{P} p(z_i = j) p(w_i|z_i = j) p(\mathbf{x}_i|z_i = j, \boldsymbol{\theta})$$

- Part 0 is the background (broad distributions for w and $\mathbf{x}$)
- $p(\mathbf{x}_i|z_i = j, \boldsymbol{\theta})$ will contain geometric information, e.g. relative offset of part j from the centre of the model

$$p(W, X|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(w_i, \mathbf{x}_i|\boldsymbol{\theta})$$

$$p(W, X) = \int p(W, X|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$



Fergus, Perona, Zisserman (2005)

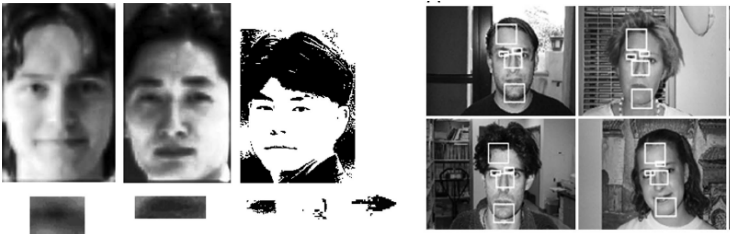
Results and Discussion

- Sudderth et al’s model is generative, and can be trained unsupervised (cf Serre et al)
- There is not much in the way of top-down influences (except rôle of $\boldsymbol{\theta}$)
- The model doesn’t do segmentation
- Use of *context* should boost performance
- There is still much to be done to obtain human level performance!

Including top-down interaction

- Extensive top-down connections everywhere in the brain
- One known role: attention. For the rest: many theories

[**?**]



Local parts can be ambiguous, but knowing global object at helps.
Top-down to set priors.
Improvement in object recognition is actually small,
but recognition and localization of parts is much better.

References I