Architectures for Visual Recognition

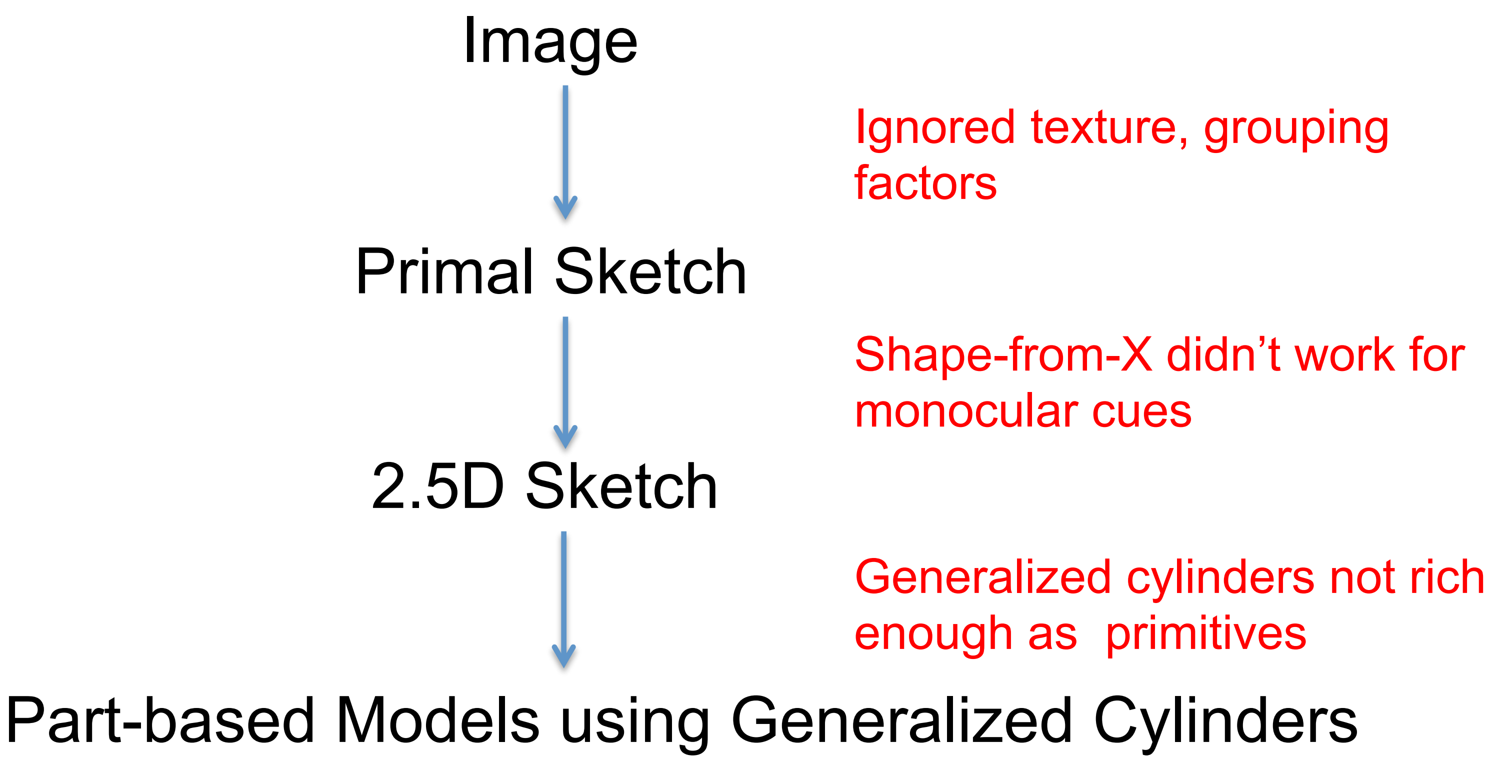
Jitendra Malik

UC Berkeley

Theories of Visual Perception in the 20th century

- Behaviorism emphasized stimulus generalization and association. Aligns well with machine learning approaches to recognition.
- Gestaltists emphasized perceptual organization- grouping and figure/ground phenomena. Natural home for those who regard reorganization of the stimulus – from pixels to entities- as primary.
- Gibson’s ecological optics emphasized “information pickup” by a moving observer. Introduced optic flow and texture gradients as powerful 3d cues. Consistent with a view that there is enough information for 3d reconstruction of the world.

Marr's paradigm (1980)



Image

Ignored texture, grouping factors

Primal Sketch

Shape-from-X didn't work for monocular cues

2.5D Sketch

Generalized cylinders not rich enough as primitives

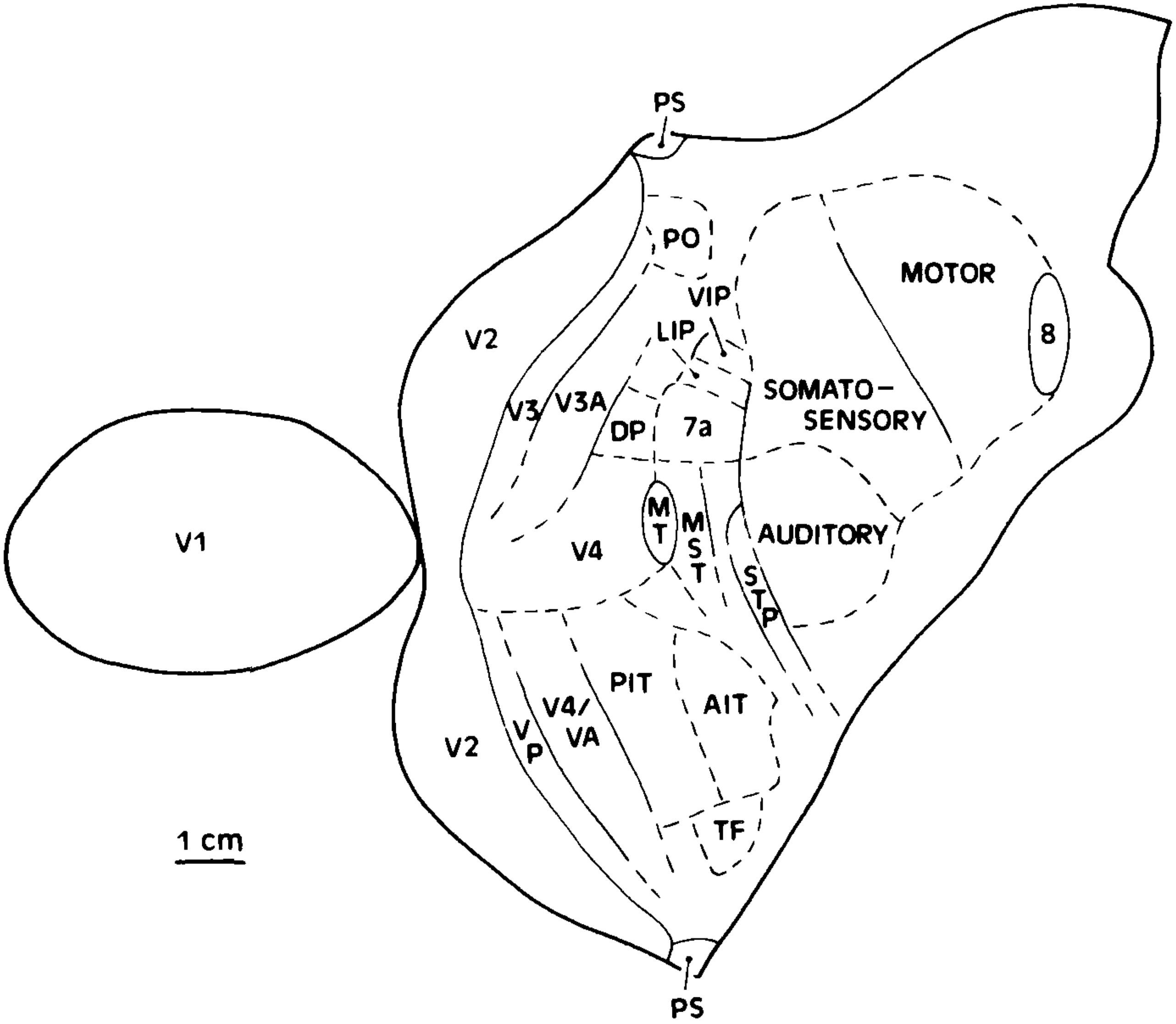
Part-based Models using Generalized Cylinders

Overall approach violated the principle of least commitment, that Marr had himself advocated. Didn't use probabilistic inference or learning.

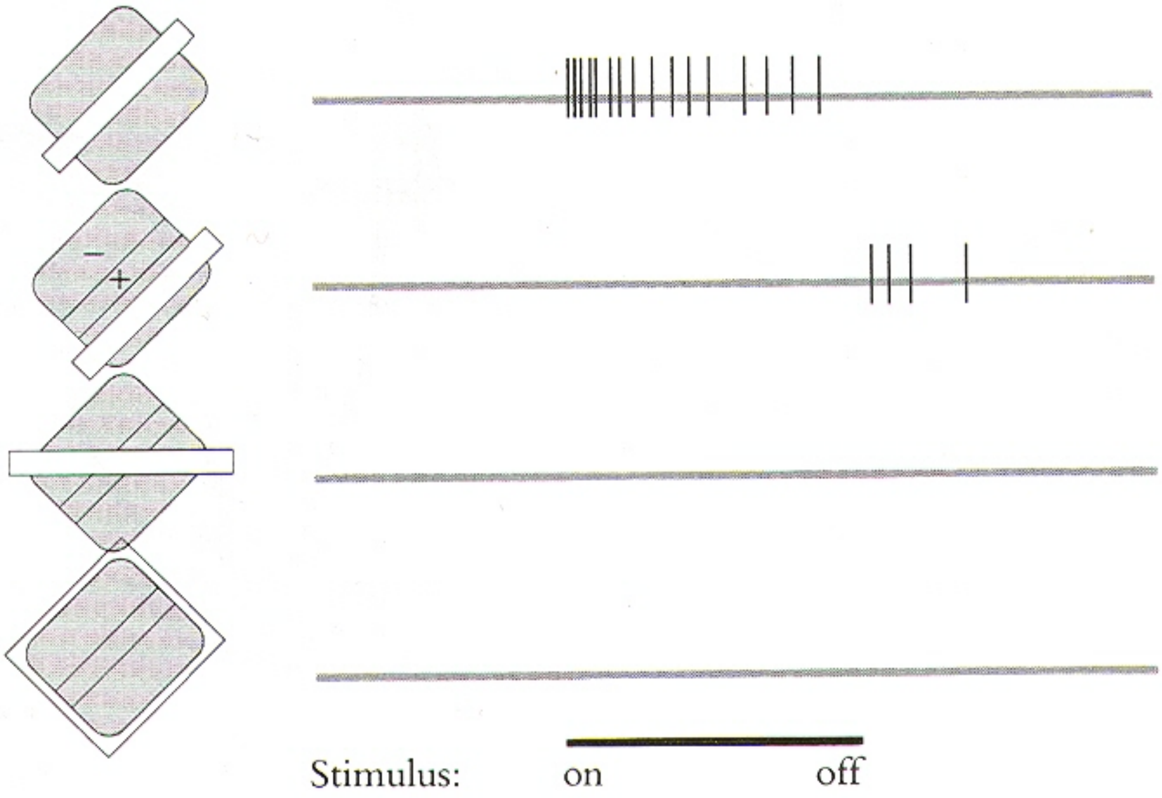
Computer vision since 1990…

- Significant progress **without** an overarching theory
- Has made considerable use of models drawn from
  - Geometry
  - Statistics/Machine learning
  - Optimization

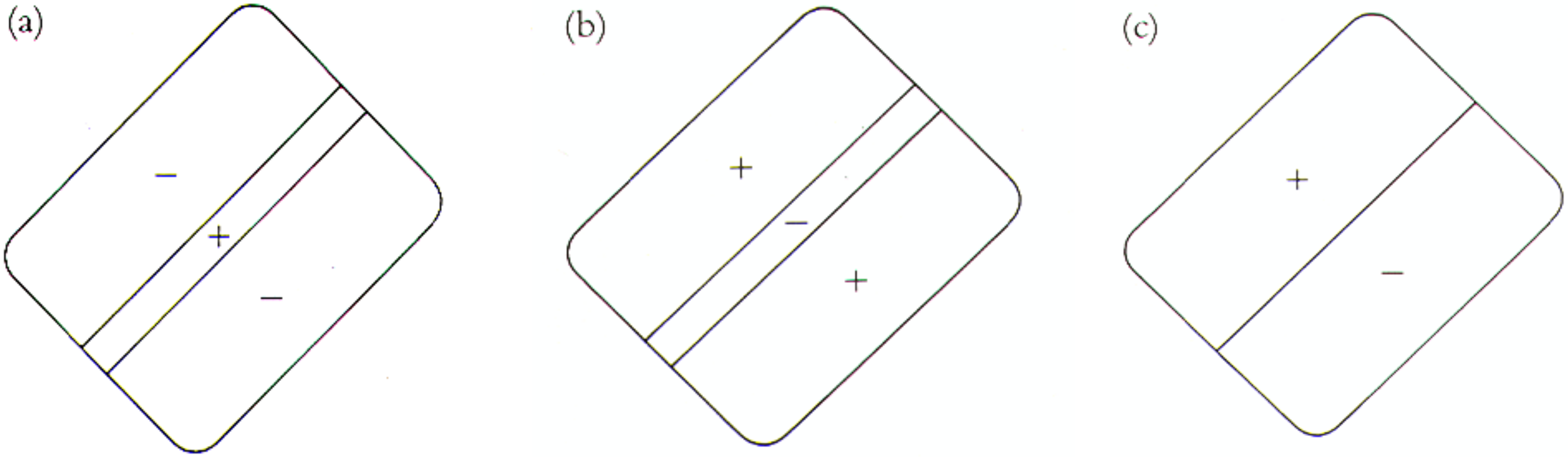
Can neuroscience guide the search for an architecture for computer vision?

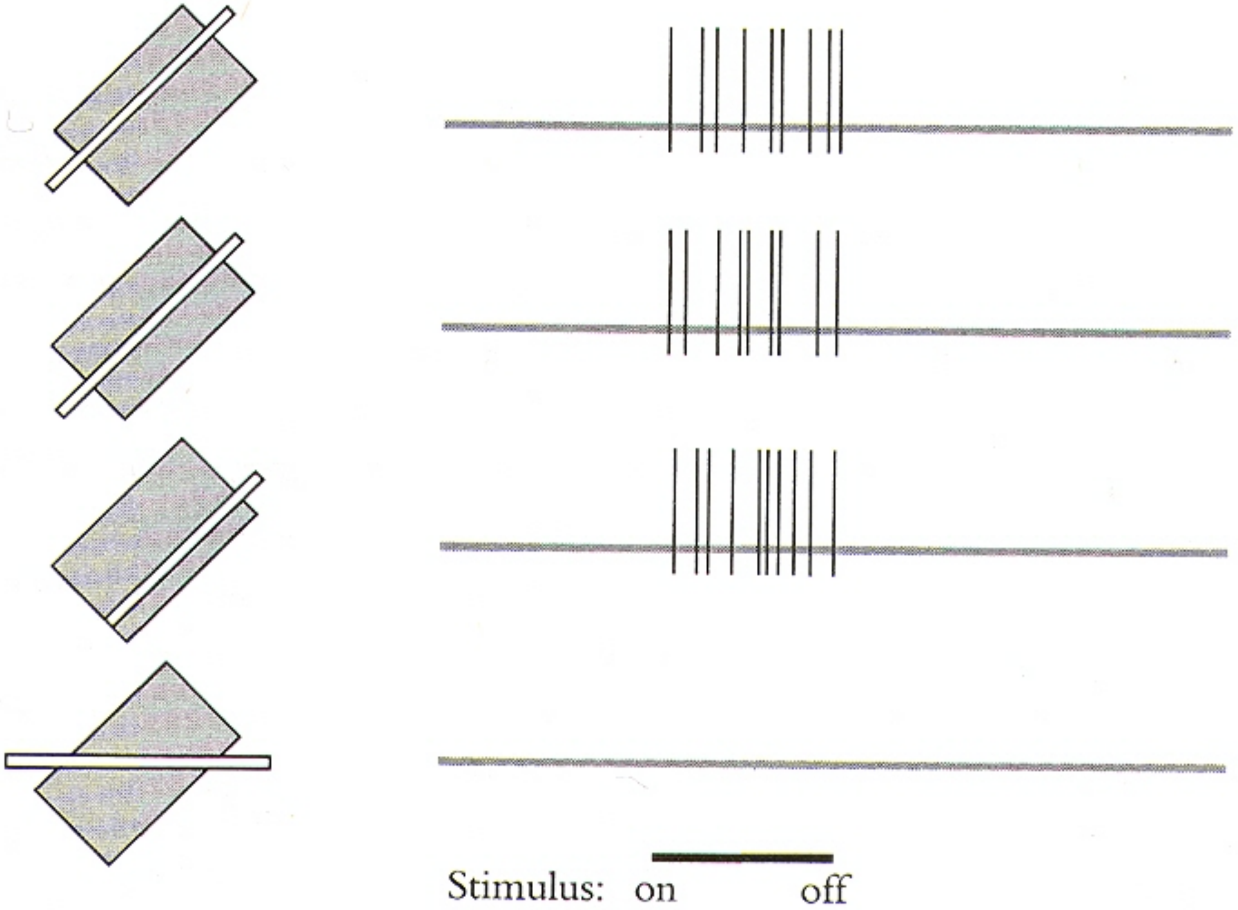


Hubel and Wiesel (1962) discovered orientation sensitive neurons in V1

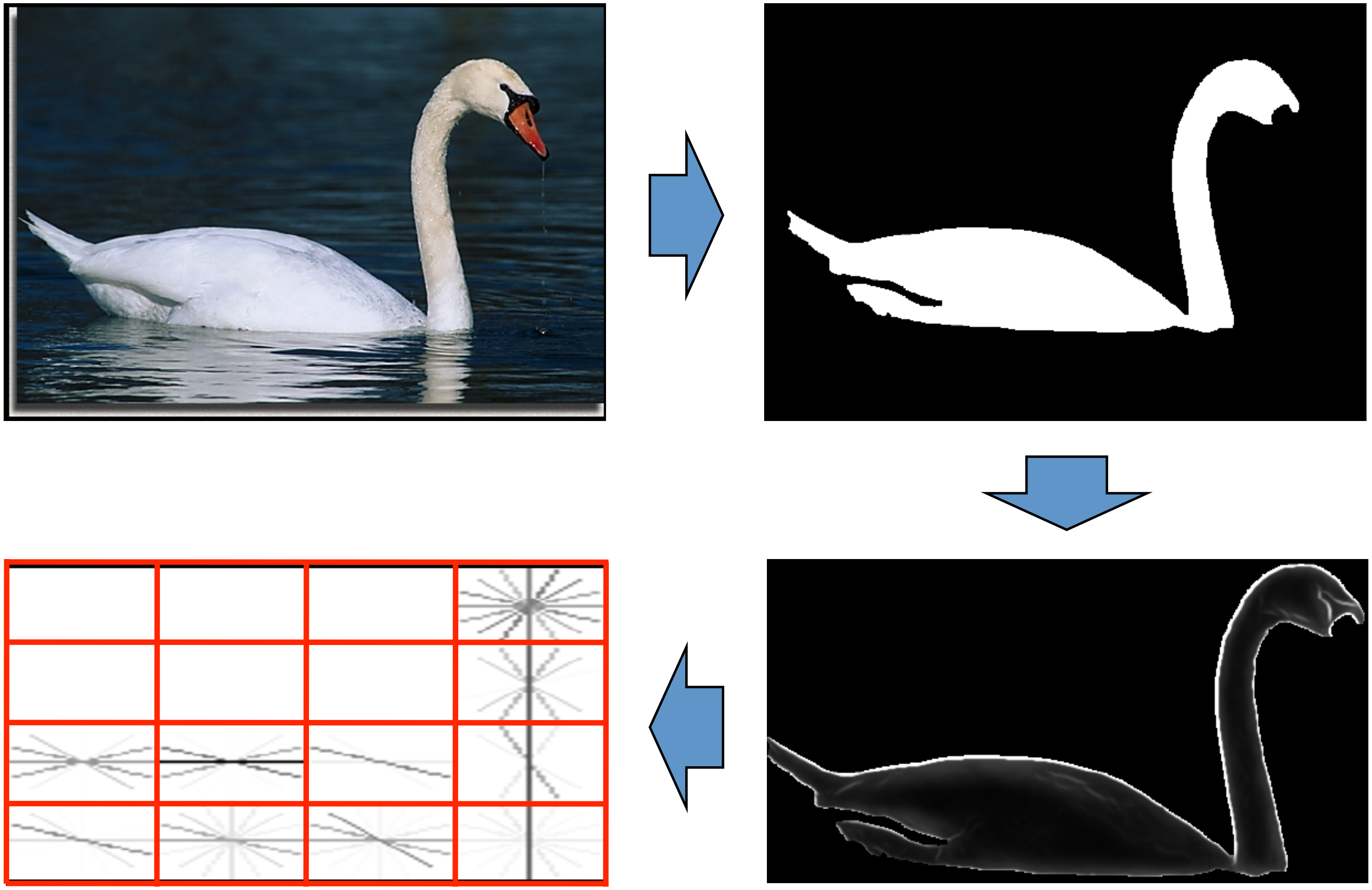


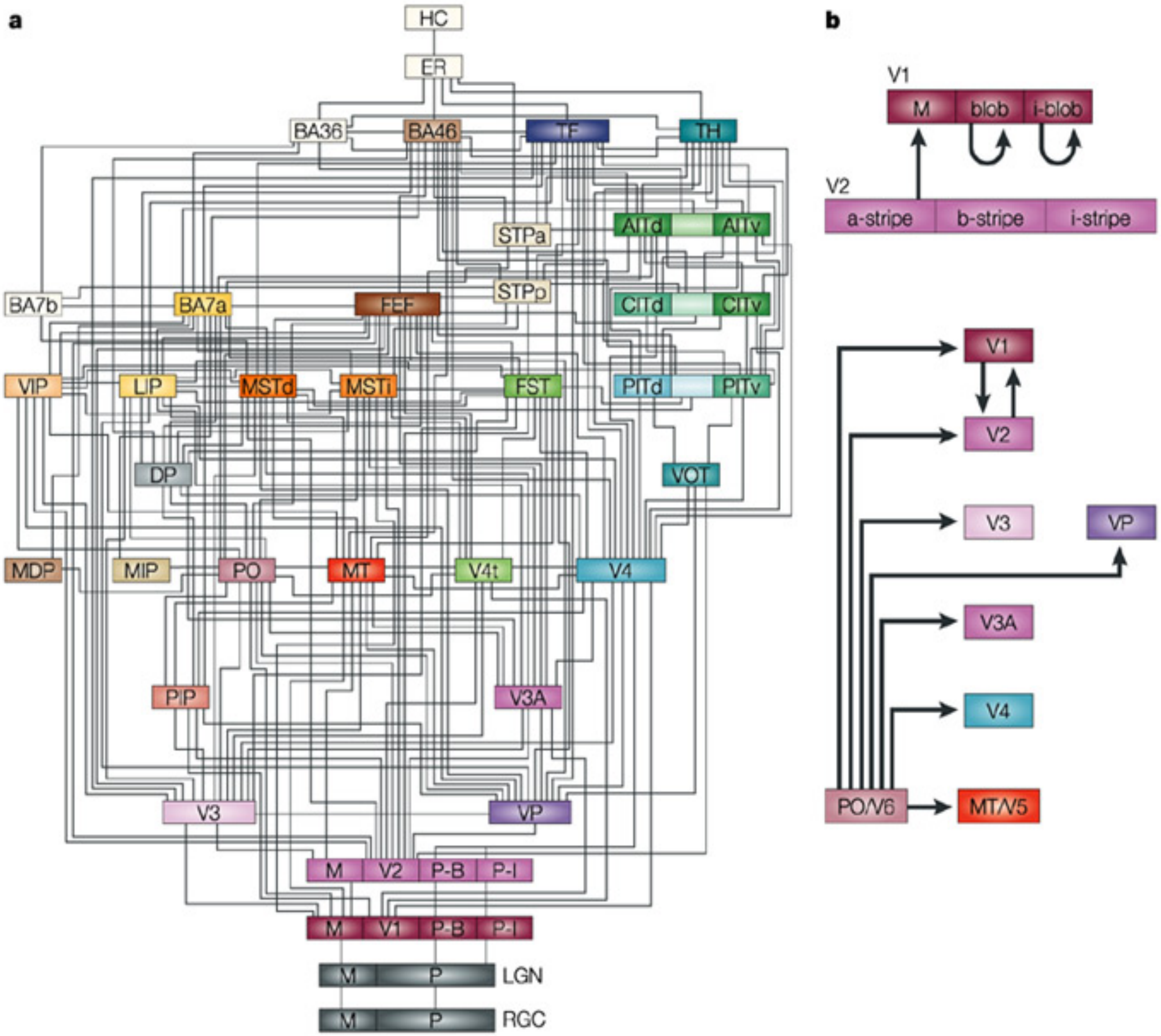
These cells respond to edges and bars ..




Stimulus: on off

Orientation based features were inspired by V1 (SIFT, GIST, HOG, GB etc)





Nature Reviews | Neuroscience

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

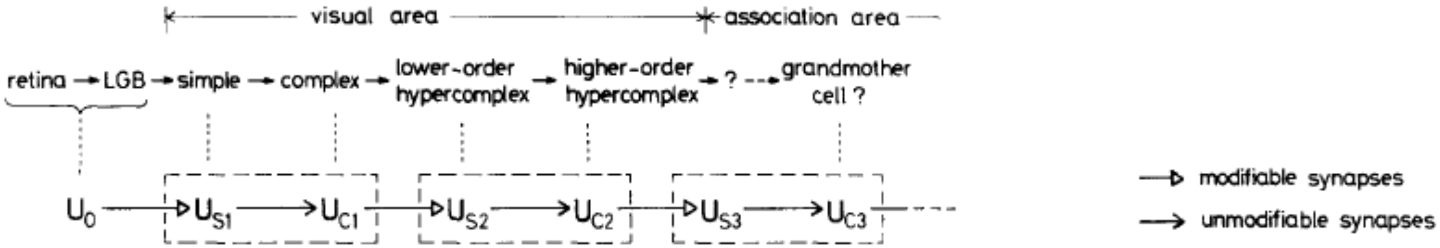


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

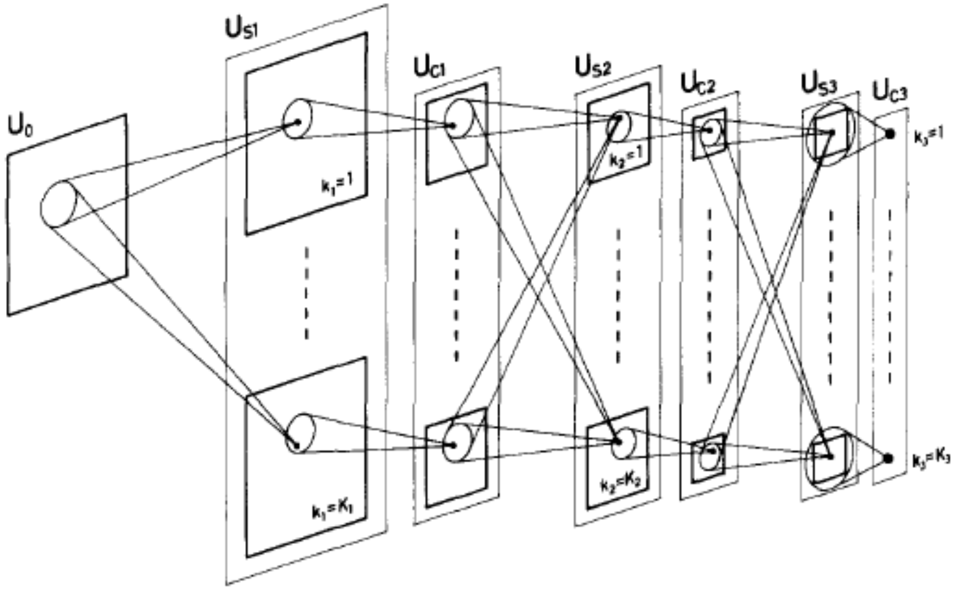
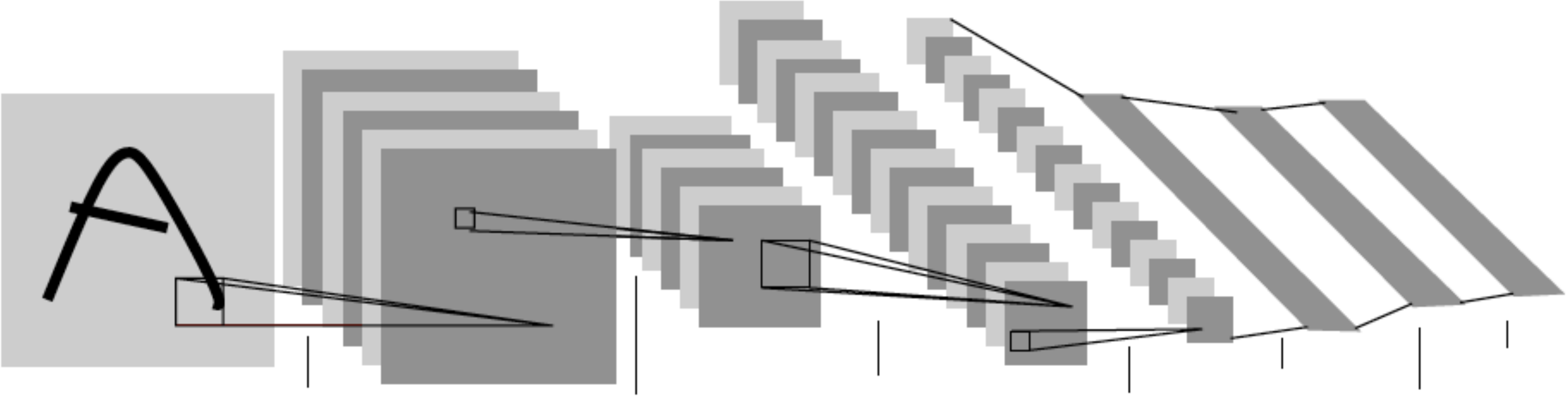


Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Biol. Cybernetics 36, 193–202 (1980)

Convolutional Neural Networks (LeCun)

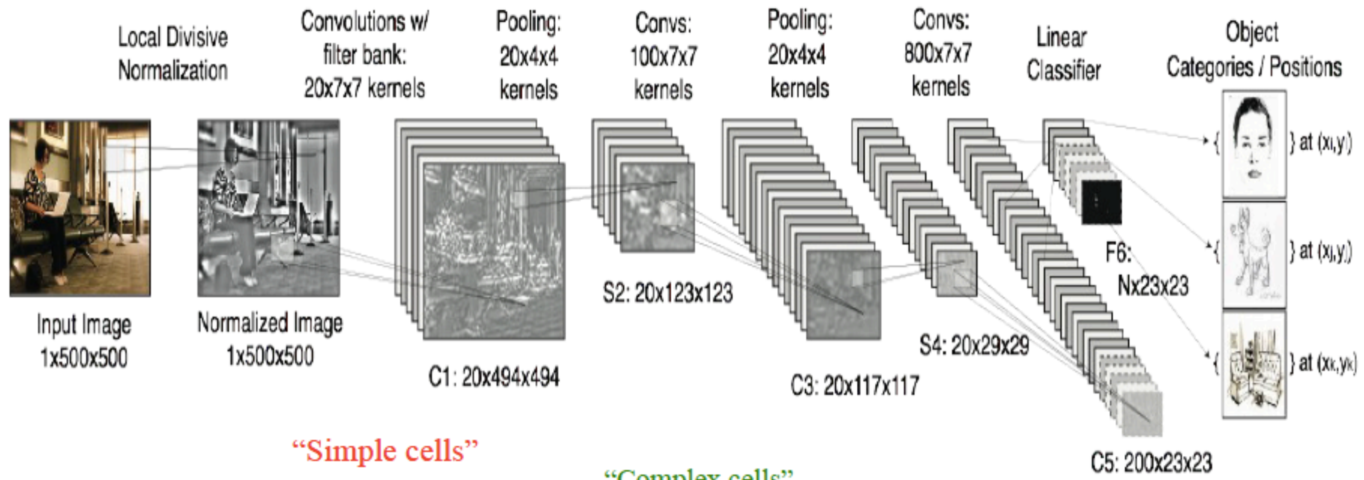
- Multilayer perceptrons with weight sharing
- LeCun showed their effectiveness for problems such as handwritten digit recognition back in the 1990s
- Recent excitement under the label of “Deep Learning”. Krizhevsky, Sutskever & Hinton (2012) showed impressive results on image classification at the ImageNet Challenge

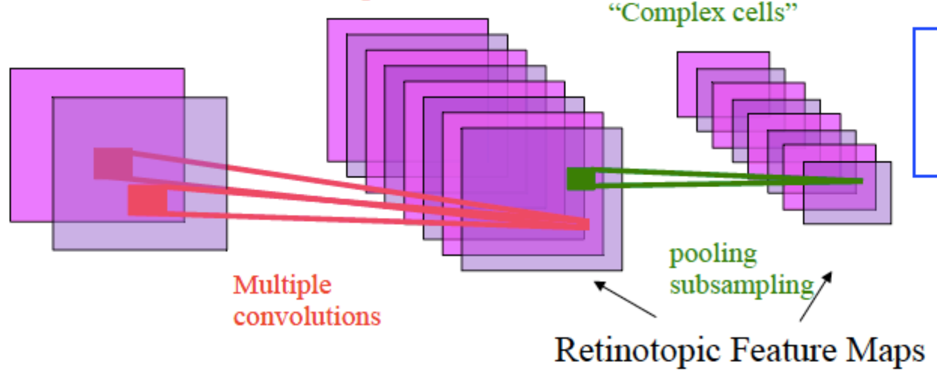


The next few slides are taken from Yann LeCun’s presentation at CVML, Paris, 2013

The Convolutional Net Model (Multistage Hubel-Wiesel system)

Y LeCun





- Training is supervised
- With stochastic gradient descent

[LeCun et al. 89]

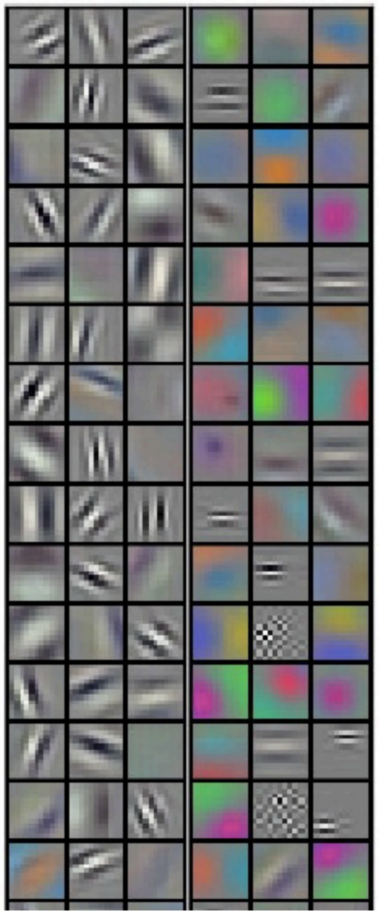
[LeCun et al. 98]

Object Recognition [Krizhevsky, Sutskever, Hinton 2012]

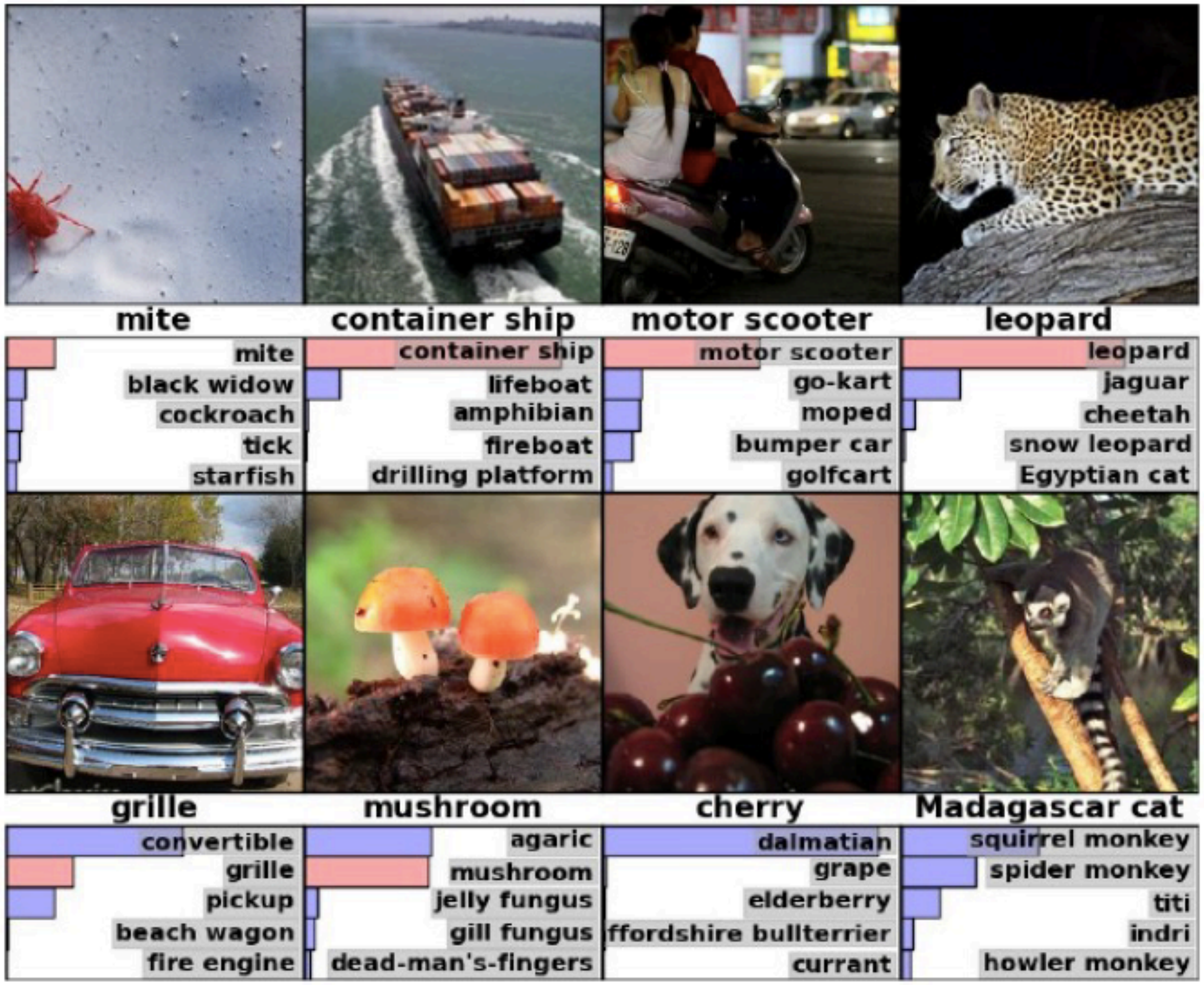
- Method: large convolutional net
  - 650K neurons, 832M synapses, 60M parameters
  - Trained with backprop on GPU
  - Trained "with all the tricks Yann came up with in the last 20 years, plus dropout" (Hinton, NIPS 2012)
  - Rectification, contrast normalization,...
- Error rate: 15% (whenever correct class isn't in top 5)
- Previous state of the art: 25% error

- A REVOLUTION IN COMPUTER VISION

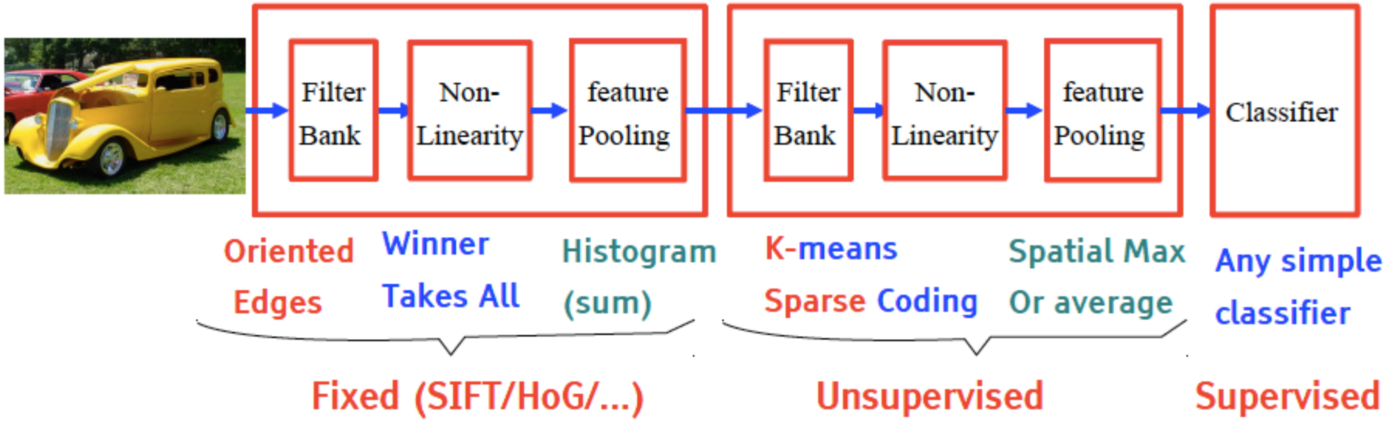
- Acquired by Google in Jan 2013
- Deployed in Google+ Photo Tagging in May 2013



Object Recognition [Krizhevsky, Sutskever, Hinton 2012]



“Mainstream” object recognition pipeline 2006-2012: somewhat similar to ConvNets



Filter Bank → Non-Linearity → feature Pooling → Filter Bank → Non-Linearity → feature Pooling → Classifier

Oriented Edges

Winner Takes All

Histogram (sum)

K-means Sparse Coding

Spatial Max Or average

Any simple classifier

Fixed (SIFT/HoG/...)

Unsupervised

Supervised

- Fixed Features + unsupervised mid-level features + simple classifier
  - SIFT + Vector Quantization + Pyramid pooling + SVM
    - [Lazebnik et al. CVPR 2006]
  - SIFT + Local Sparse Coding Macrofeatures + Pyramid pooling + SVM
    - [Boureau et al. ICCV 2011]
  - SIFT + Fisher Vectors + Deformable Parts Pooling + SVM
    - [Perronin et al. 2012]

My opinion…

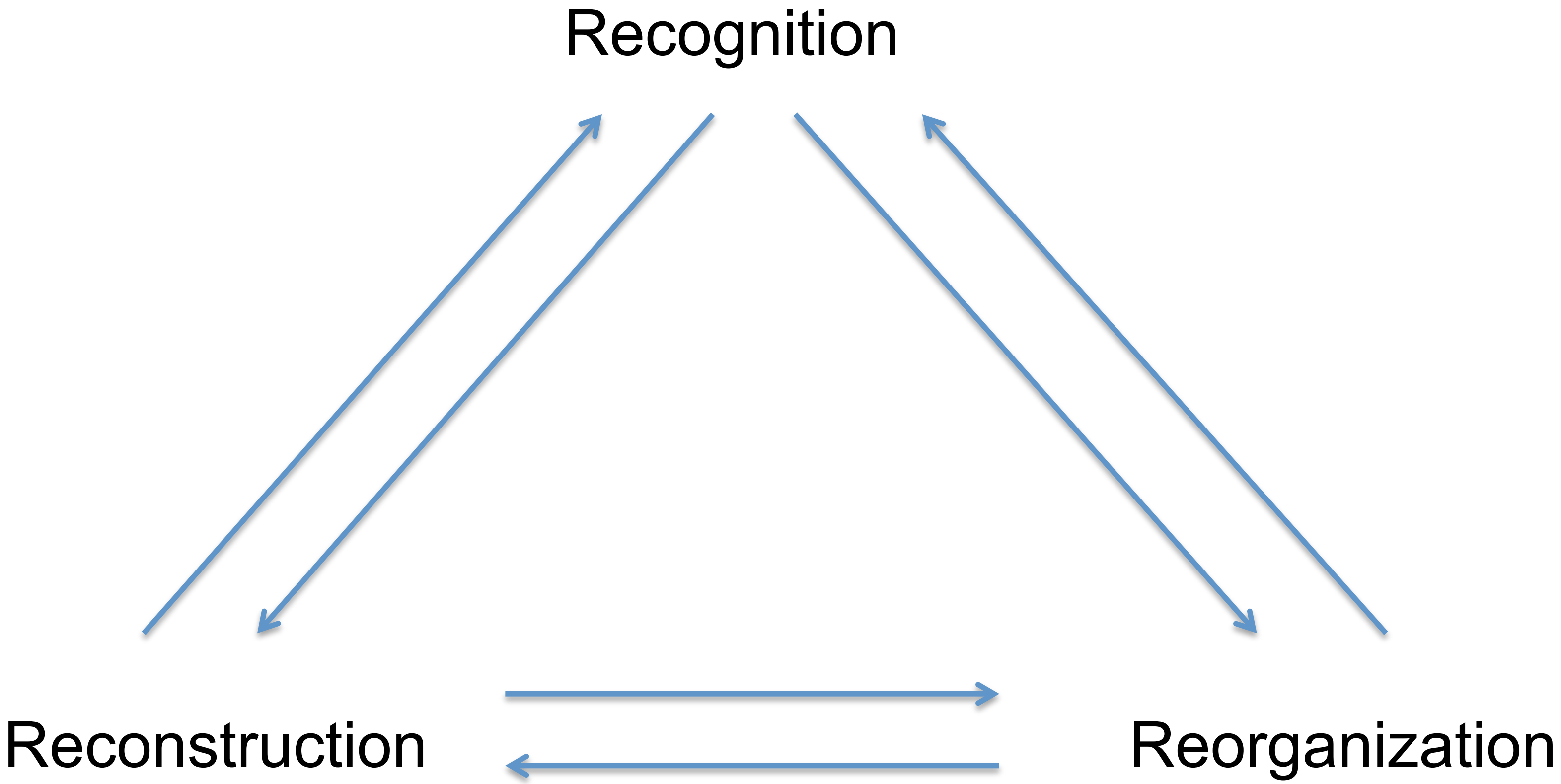
- The availability of “big data” means that high capacity learning machines have greater potential than before. We can choose a different point on the bias-variance tradeoff from machine learning theory.
- Multilayer neural networks are not the only way. For example, random forests, with suitably rich set of questions, could do so as well.
- The experience of handwritten digit recognition where 5 or 6 different approaches achieve below 1% error rates suggests that it is not worth having a religious battle over classifiers.
- But, there is more to vision than classification!

Different aspects of vision

- Perception: study the “laws of seeing” -predict what a human would perceive in an image.
- Neuroscience: understand the mechanisms in the retina and the brain
- Function: how laws of optics, and the statistics of the world we live in, make certain interpretations of an image more likely to be valid

The match between human and computer vision is strongest at the level of function, but since typically the results of computer vision are meant to be conveyed to humans makes it useful to be consistent with human perception. Neuroscience is a source of ideas but being bio-mimetic is not a requirement.

The Three R’s of Vision



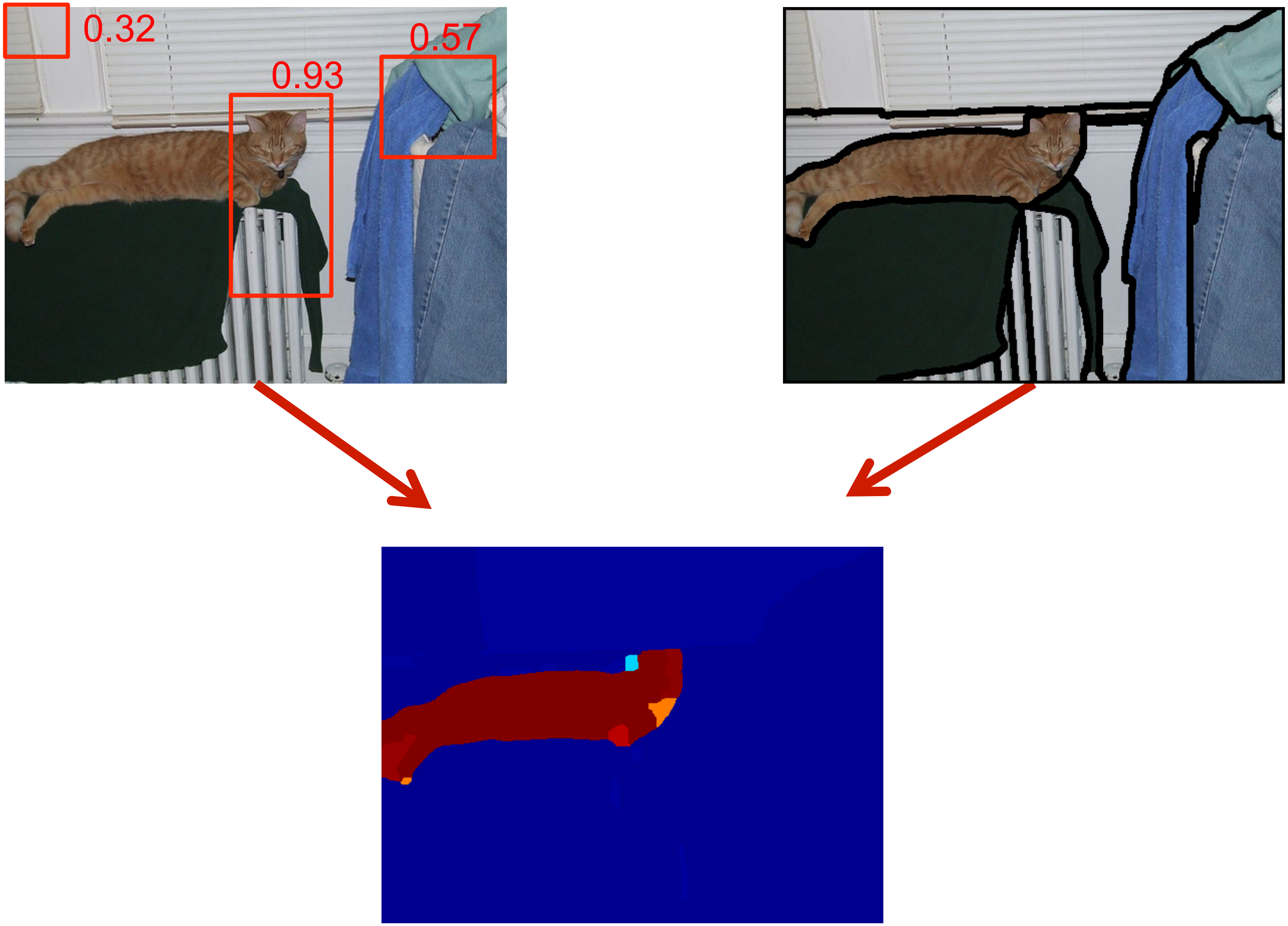
Recognition

Reconstruction

Reorganization

Each of the 6 directed arcs in this diagram is a useful direction of information flow

Why templates need to be combined with regions



Black and White Tights Dance

- Demo from youtube

Some remarks..

- Child development studies clearly show the importance of grouping based on common motion. This capability is present very early and helps train other grouping cues.
- This gives a natural mechanism for “objectness”; tracking over time trains visual correspondence.
- In computer vision, I have long argued for using bottom-up segmentation as a way to generate candidates for recognition. But it is still the case that these approaches don’t do as well as “naïve” sliding window approaches; inevitably some recall is lost.
- Video analysis should work much better than static image analysis for this purpose; it is a pity that so little work has been done on this (see Brox for a counter example)
- Read the last section of Wertheimer (1923) for an incisive discussion on the need for perceptual organization

PSYCHOLOGISCHE FORSCHUNG

ZEITSCHRIFT FÜR PSYCHOLOGIE
UND IHRE GRENZWISSENSCHAFTEN

HERAUSGEGEBEN VON

K. KOFFKA
GIESSEN

W. KÖHLER
BERLIN

M. WERTHEIMER
BERLIN

K. GOLDSTEIN
FRANKFURT A. M.

H. GRUHLE
HEIDELBERG

VIERTER BAND
MIT 125 TEXTABBILDUNGEN

FESTSCHRIFT FÜR CARL STUMPF

BERLIN
VERLAG VON JULIUS SPRINGER
1923

Untersuchungen zur Lehre von der Gestalt.

II.

Von
Max Wertheimer.

Mit 56 Abbildungen im Text.

Ich stehe am Fenster und sehe ein Haus, Bäume, Himmel.

Und könnte nun, aus theoretischen Gründen, abzuzählen versuchen und sagen: da sind ... 327 Helligkeiten (und Farbtöne).

(Habe ich „327"? Nein; Himmel, Haus, Bäume; und das Haben der „327" als solcher kann keiner realisieren.)

Und seien in dieser sonderbaren Rechnung etwa Haus 120 und Bäume 90 und Himmel 117, so habe ich jedenfalls *dieses* Zusammen, dieses Getrenntsein, und nicht etwa 127 und 100 und 100; oder 150 und 177.

In dem bestimmten Zusammen, der bestimmten Getrenntheit *sehe* ich es; und in welcher Art des Zusammen, der Getrenntheit ich es sehe, das steht nicht einfach in meinem Belieben: ich kann durchaus nicht etwa nach Belieben jede irgend andere gewünschte Art der Zusammengefaßtheit einfach realisieren.

(Und welch ein merkwürdiger Prozeß, wenn einmal so etwas gelingt. Welches Erstaunen, wie ich hier nach langem Hinsehen, nach allerlei Versuchen, in sehr wirklichkeitsferner Einstellung *entdeckte*, daß da an einem Fenster Stücke des dunkeln Rahmens mit einem glatten Ast zusammen ein lateinisches N bilden.) —

Oder: Die zwei Gesichter Wange an Wange. Ich sehe das eine (mit seinen, wenn man so will, „57" Helligkeiten) und das andere (mit seinen „49"); nicht aber in der Teilung 66 plus 40 oder 6 plus 100.

Theorien, die etwa fordern würden, daß ich da „106" sehe, stehen auf dem Papier; zwei Gesichter sehe ich. Aber hier mag es vorerst *nur* auf die Art des Zusammen und der Geteiltheit ankommen; die ist jedenfalls *so* bestimmt. Nur von diesem — bescheidenen, theoretisch aber nicht unwichtigen — Sachverhalt soll hier zunächst gehandelt werden.

Oder: Ich höre eine Melodie (17 Töne!) mit ihrer Begleitung (32 Töne!). Ich höre Melodie und Begleitung, nicht einfach „49" oder wenigstens gewiß nicht normaliter oder ganz nach Belieben 20 plus 29.

So ist es auch noch, wenn keinerlei Reizkontinua in Frage kommen; wenn die Melodie mit ihrer Begleitung etwa von einer der alten Spiel-

Thank you!