


# Challenges in Visual Recognition: A Historical Perspective

Jitendra Malik

University of California at Berkeley

# The more you look, the more you see!

Image shown to subjects	40ms	80ms	107ms	500ms
	<p>“Possibly outdoor scene, maybe a farm. I could not tell for sure.”</p>	<p>“There seem to be two people in the center of the scene.”</p>	<p>“ People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight.”</p>	<p>“Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets...”</p>
<p>Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].</p>				

# PASCAL Visual Object Challenge

## Dining Table



## Dog



## Horse



## Motorbike



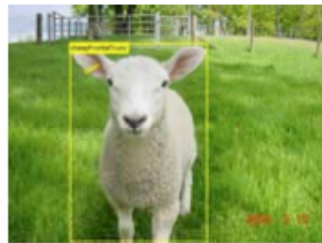
## Person



## Potted Plant



## Sheep



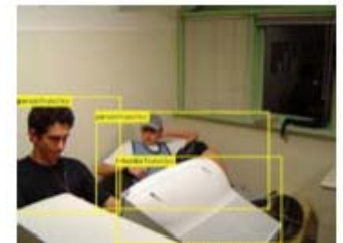
## Sofa



## Train



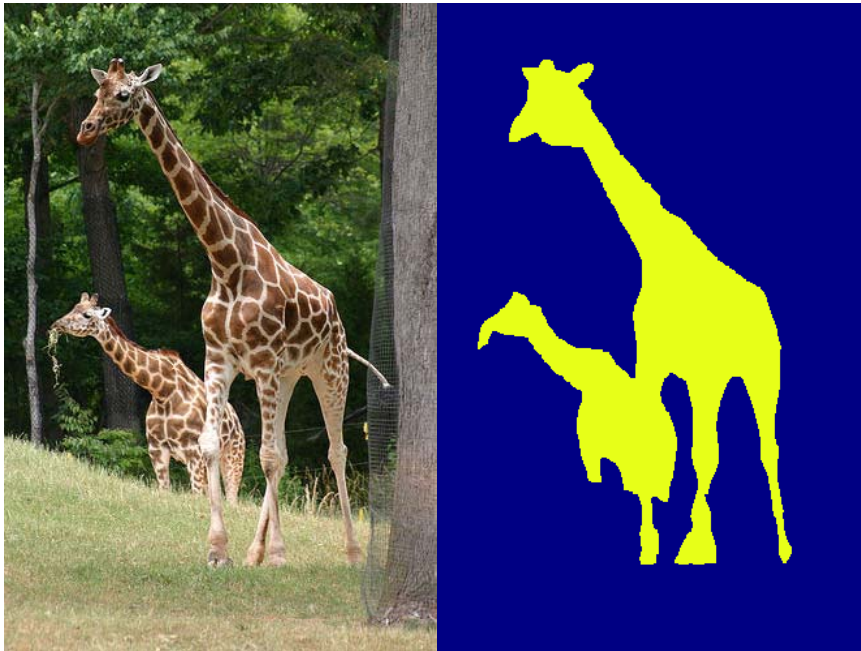
## TV/Monitor



# We want to locate the object

Orig. Image

Segmentation

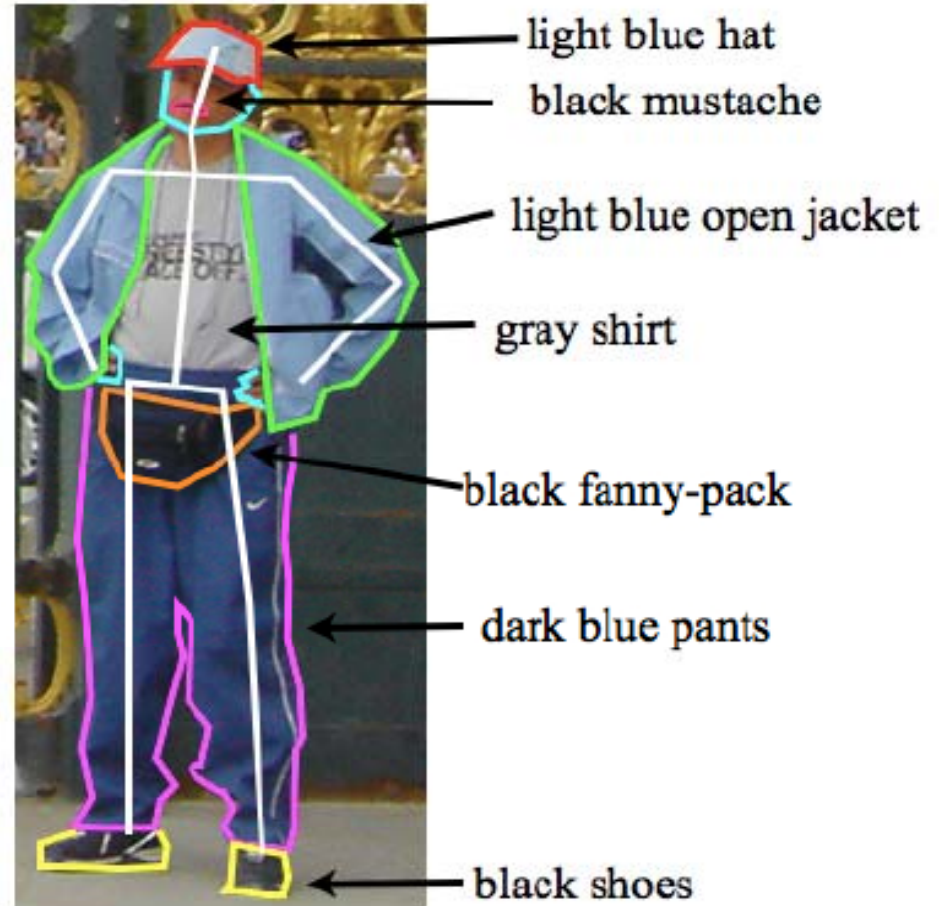
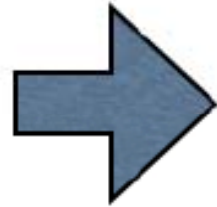


Orig. Image

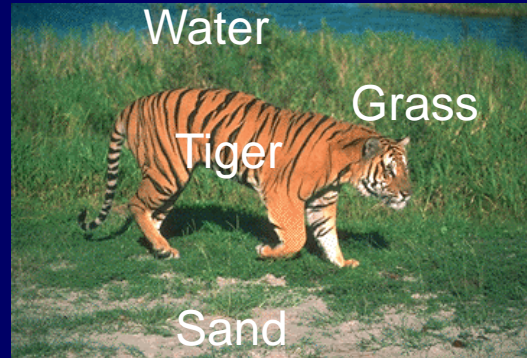
Segmentation



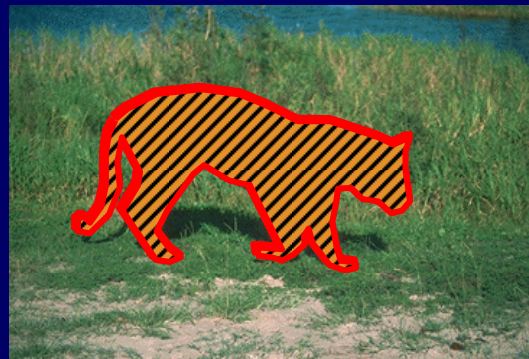
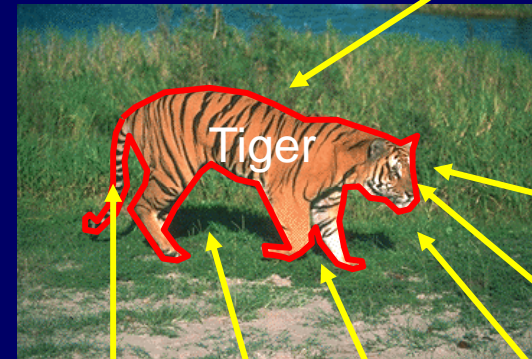
And we want to detect and label parts..



# Categorization at Multiple Levels



outdoor  
wildlife



# SUN Database: Large-scale Scene Recognition from Abbey to Zoo

Jianxiong Xiao

James Hays<sup>†</sup>

Krista A. Ehinger

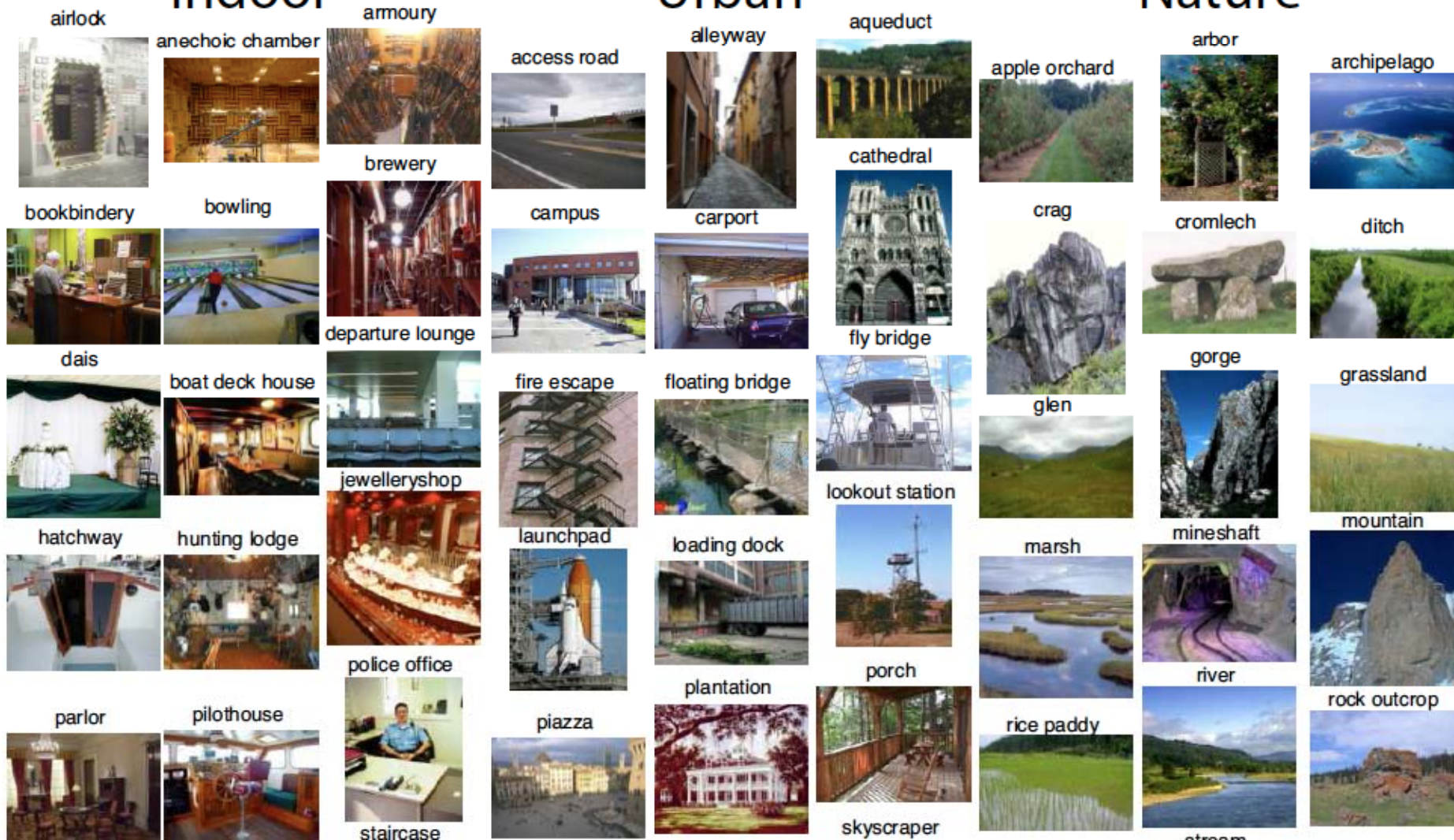
Aude Oliva

Antonio Torralba

## Indoor

## Urban

## Nature



# Examples of Actions

- **Movement and posture change**
  - run, walk, crawl, jump, hop, swim, skate, sit, stand, kneel, lie, dance (various), ...
- **Object manipulation**
  - pick, carry, hold, lift, throw, catch, push, pull, write, type, touch, hit, press, stroke, shake, stir, turn, eat, drink, cut, stab, kick, point, drive, bike, insert, extract, juggle, play musical instrument (various)...
- **Conversational gesture**
  - point, ...
- **Sign Language**



# Key cues for action recognition

- “Morpho-kinetics” of action (shape and movement of the body)
- Identity of the object/s
- Activity context
- **ACTION = MOVEMENT + GOAL**

# Resolution Regimes

## Far field



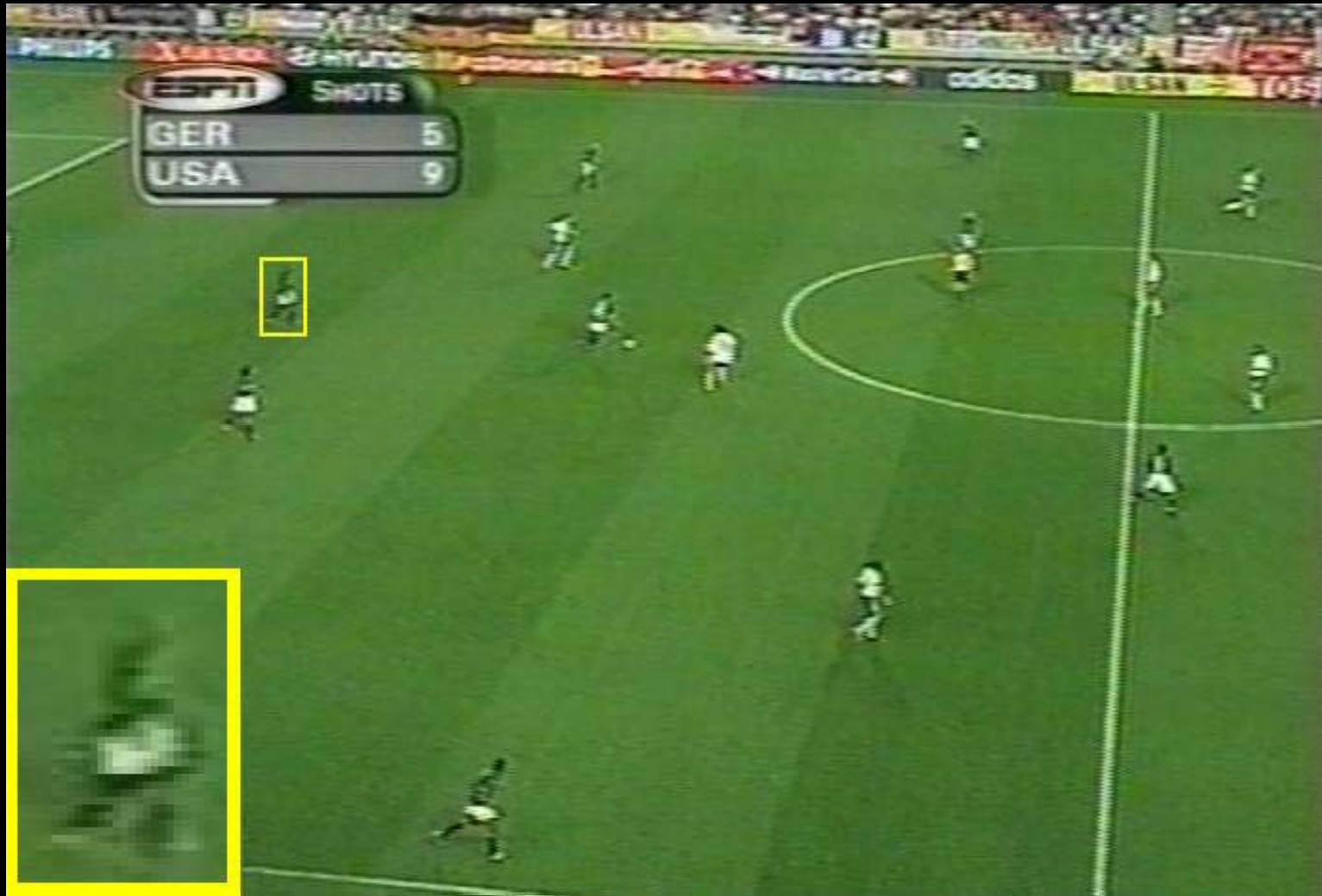
- 3-pixel man
- Blob tracking

## Near field




- 300-pixel man
- Stick Figure

# Medium-field Recognition



The 30-Pixel Man

# The more you look, the more you see!

Image shown to subjects	40ms	80ms	107ms	500ms
	<p>“Possibly outdoor scene, maybe a farm. I could not tell for sure.”</p>	<p>“There seem to be two people in the center of the scene.”</p>	<p>“ People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight.”</p>	<p>“Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets...”</p>
<p>Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].</p>				

# We need to identify

- Objects
- Agents
- Relationships among objects with objects, objects with agents, agents with agents ...
- Events and Actions

# Different aspects of vision

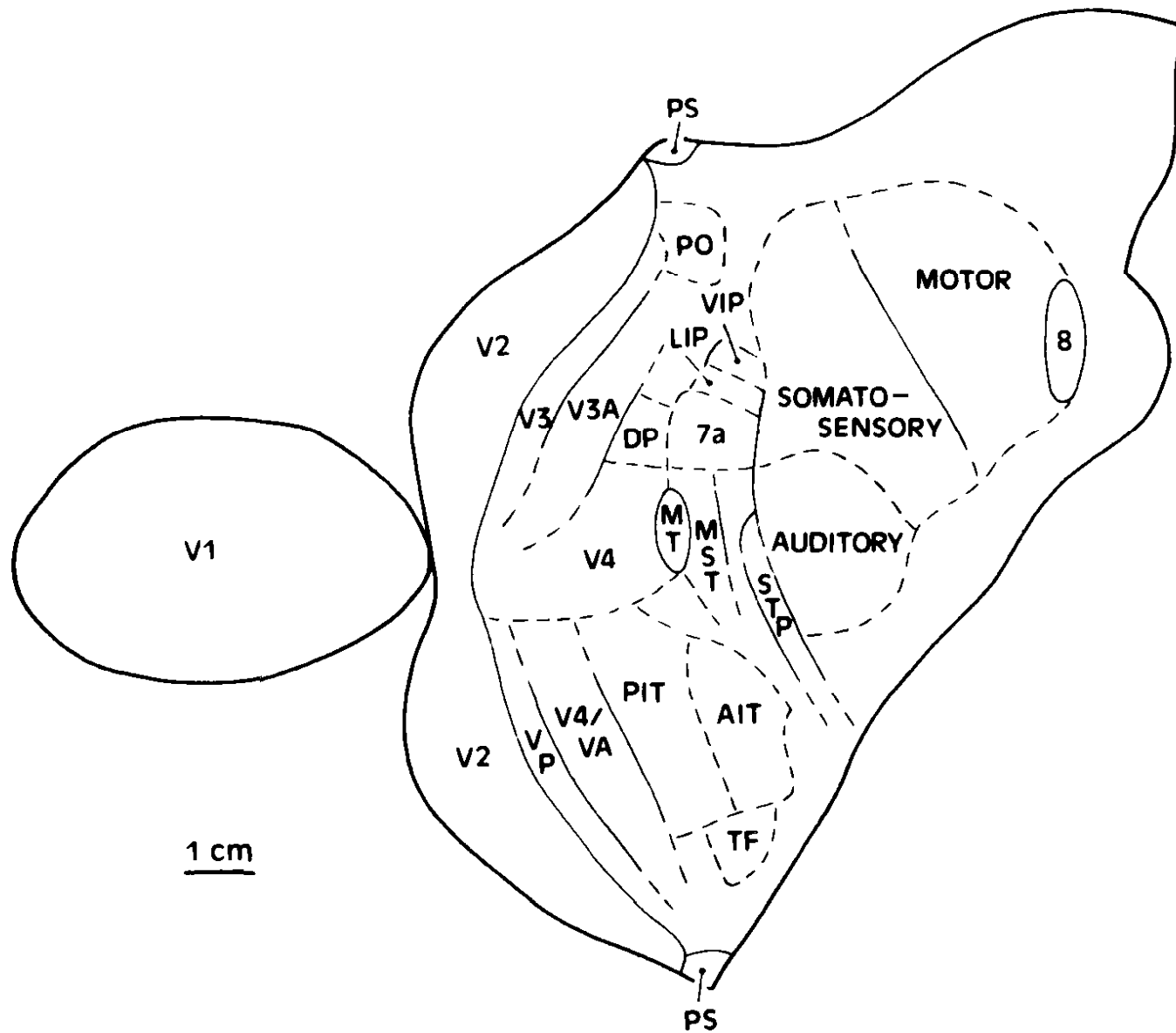
- Perception: study the “laws of seeing” -predict what a human would perceive in an image.
- Neuroscience: understand the mechanisms in the retina and the brain
- Function: how laws of optics, and the statistics of the world we live in, make certain interpretations of an image more likely to be valid

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.

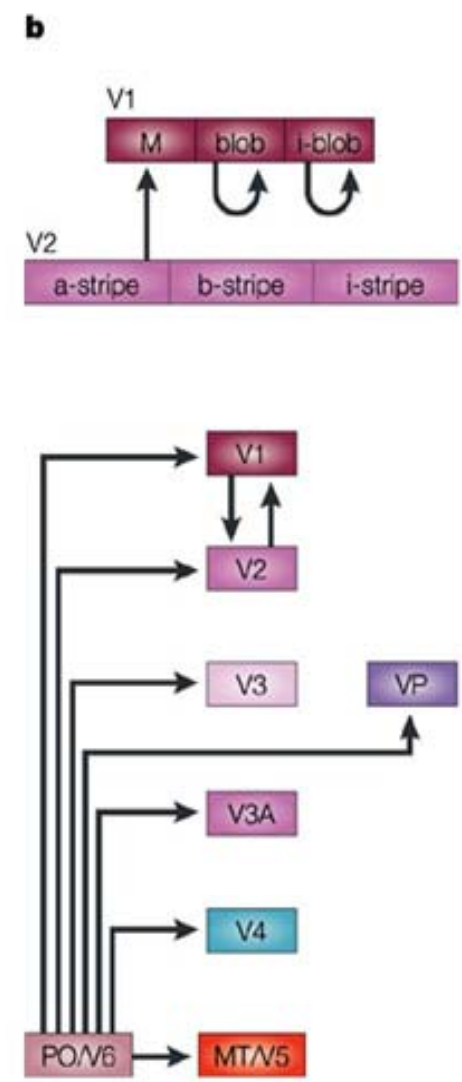
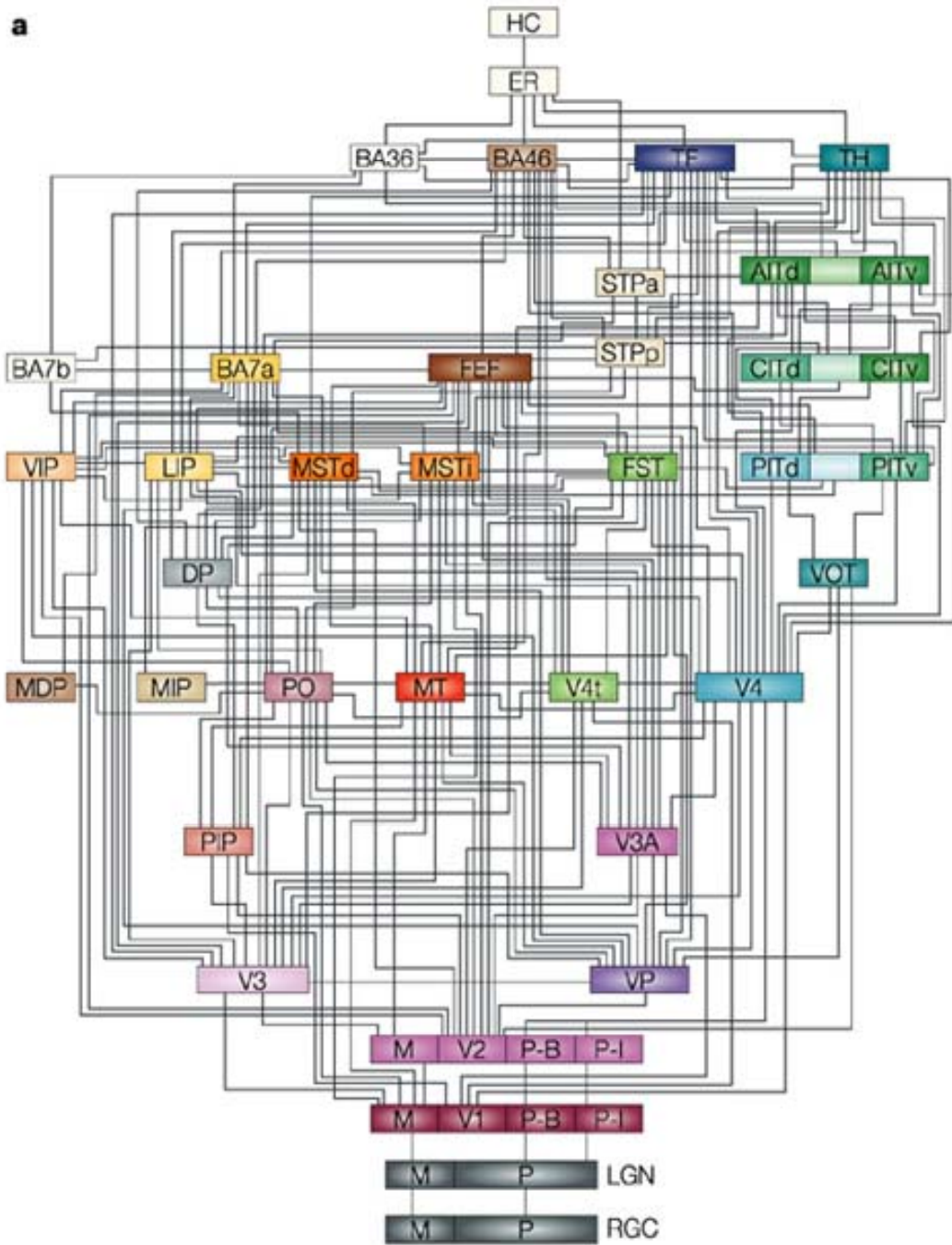
# Taxonomy and Partonomy

- Taxonomy: E.g. Cats are in the order Felidae which in turn is in the class Mammalia
  - Recognition can be at multiple levels of categorization, or be identification at the level of specific individuals , as in faces.
- Partonomy: Objects have parts, they have subparts and so on. The human body contains the head, which in turn contains the eyes.
- These notions apply equally well to scenes and to activities.
- Psychologists have argued that there is a “basic-level” at which categorization is fastest (Eleanor Rosch et al).
- In a partonomy each level contributes useful information for recognition.

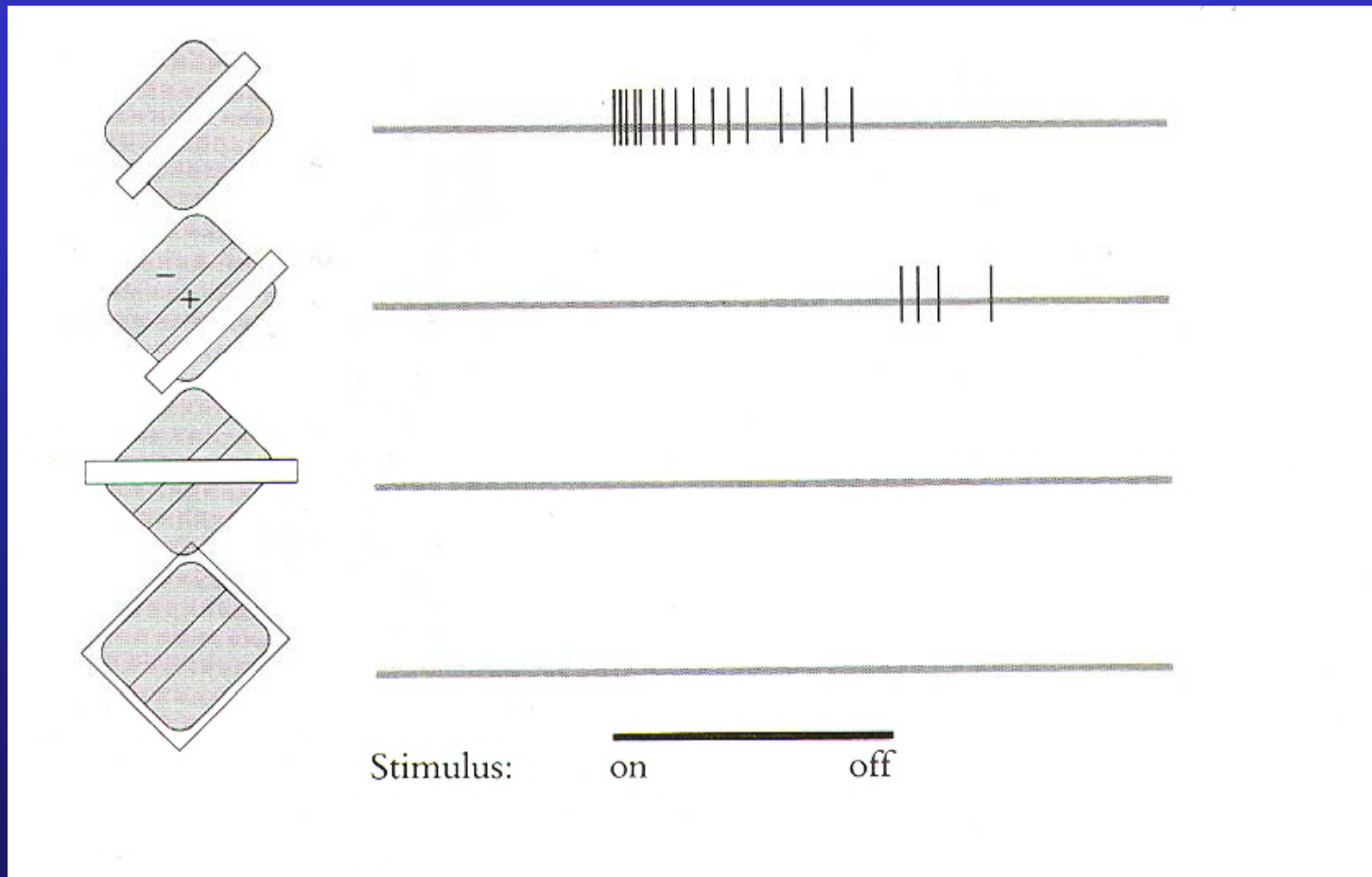
# Visual Processing Areas



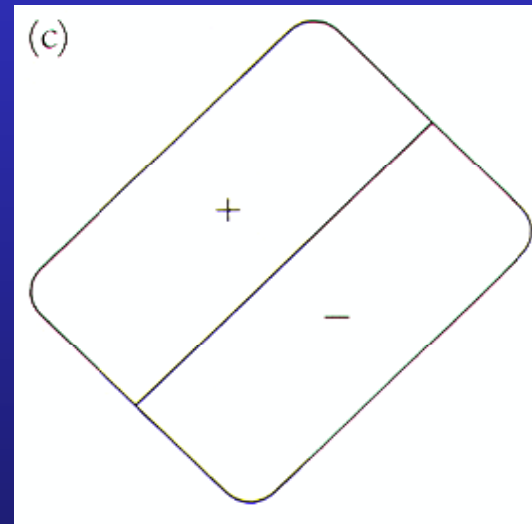
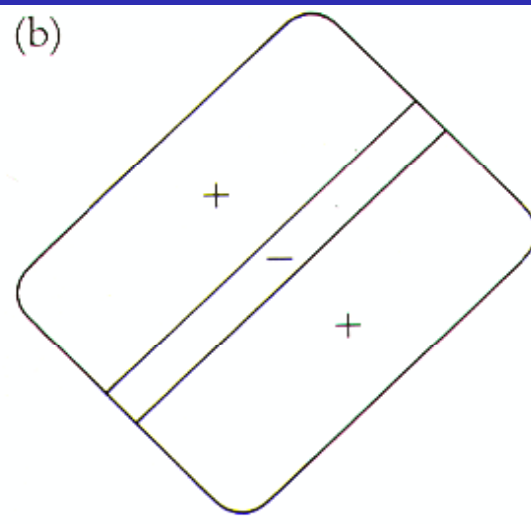
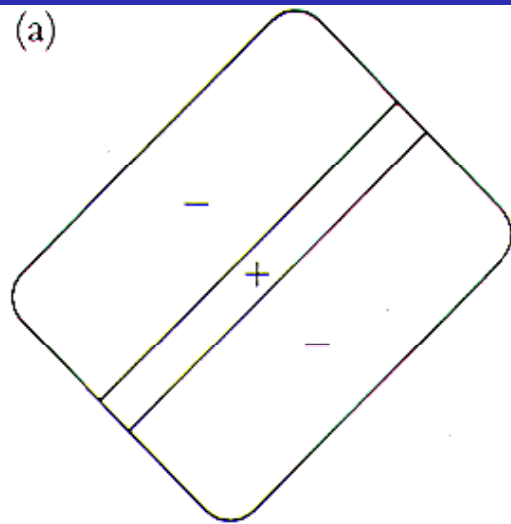


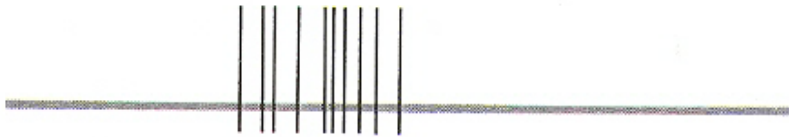
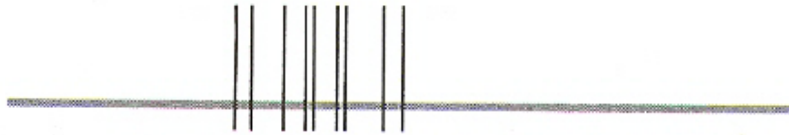
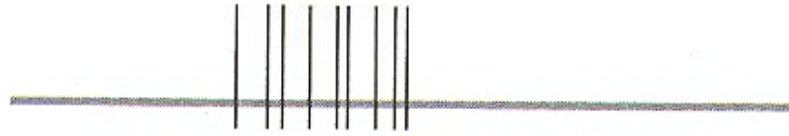
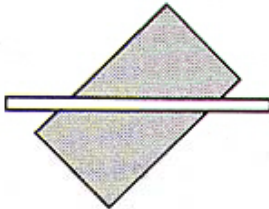
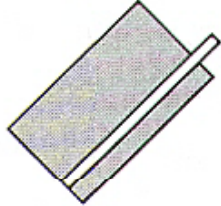
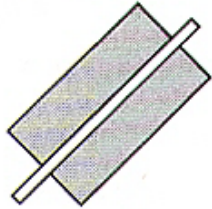
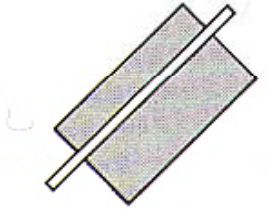



# Hubel and Wiesel (1962) discovered orientation sensitive neurons in V1



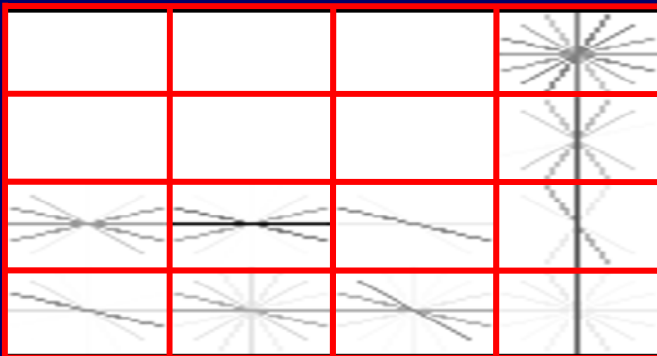
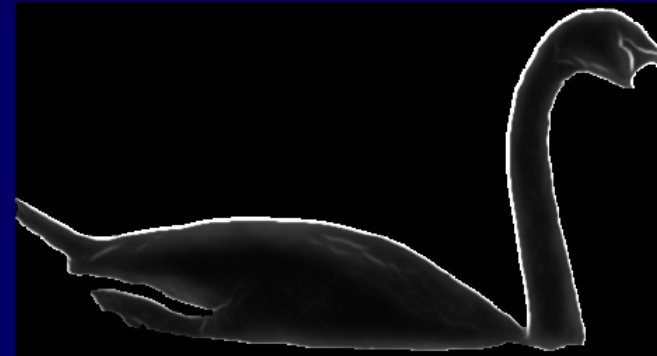
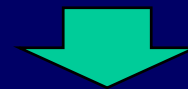
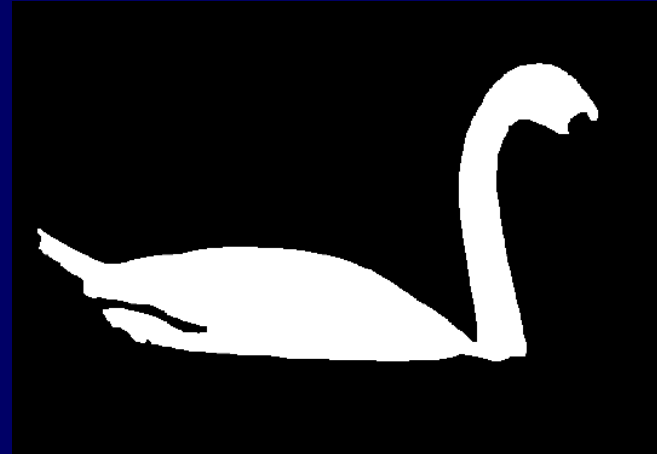
These cells respond to edges and bars ..





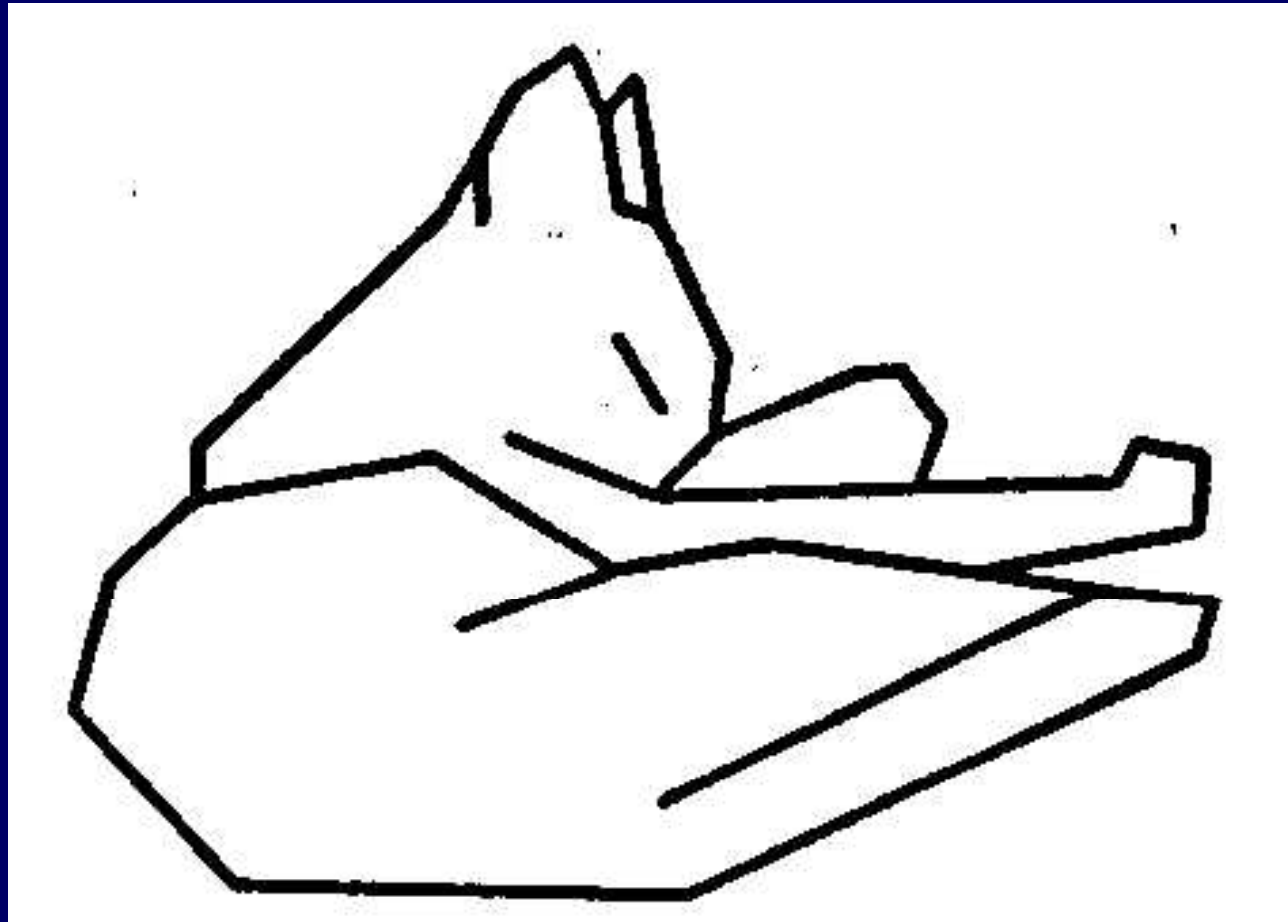
Stimulus:  on off

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



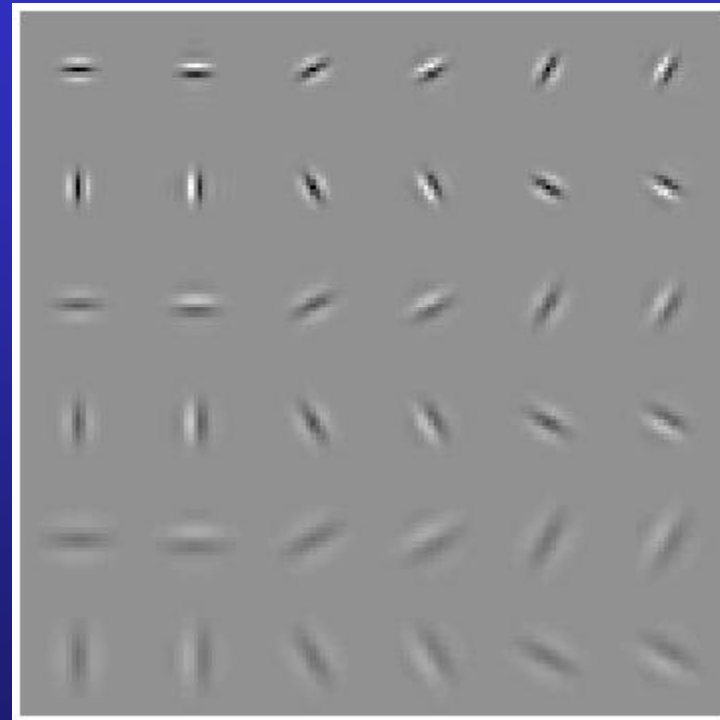
# Attneave's Cat (1954)

**Line drawings convey most of the information**



# Modeling simple cells

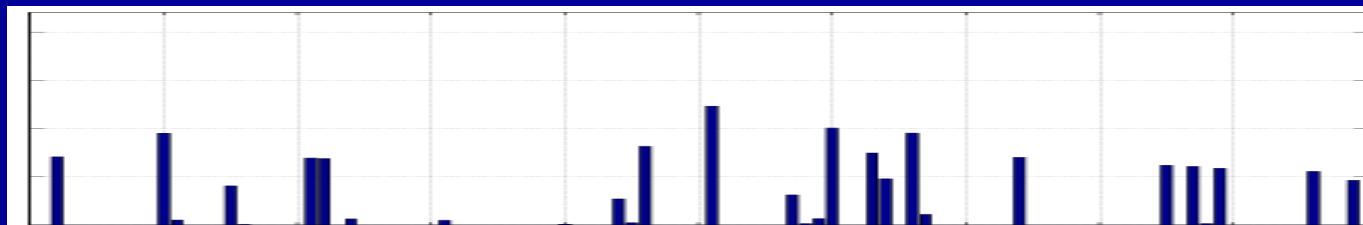
- Elongated directional Gaussian derivatives
- 2nd derivative and Hilbert transform
- $L_1$  normalized for scale invariance
- 6 orientations, 3 scales
- Zero mean



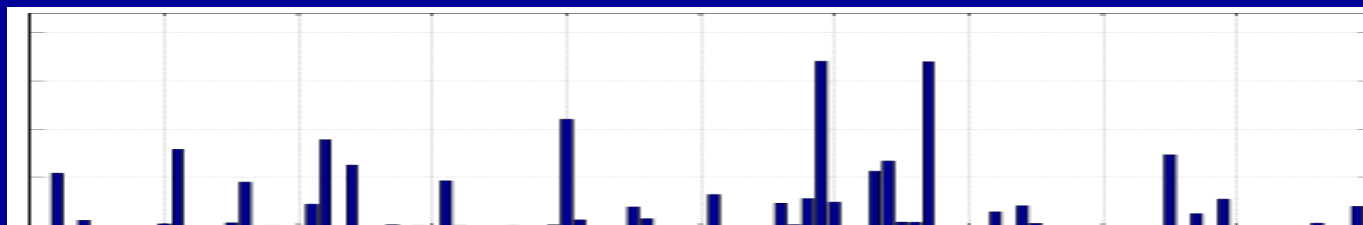
Used for texture discrimination and classification by Malik and Perona (1990), Leung and Malik (1999)

# Texton Histogram Model for Recognition (Leung & Malik, 1999) cf. Bag of Words

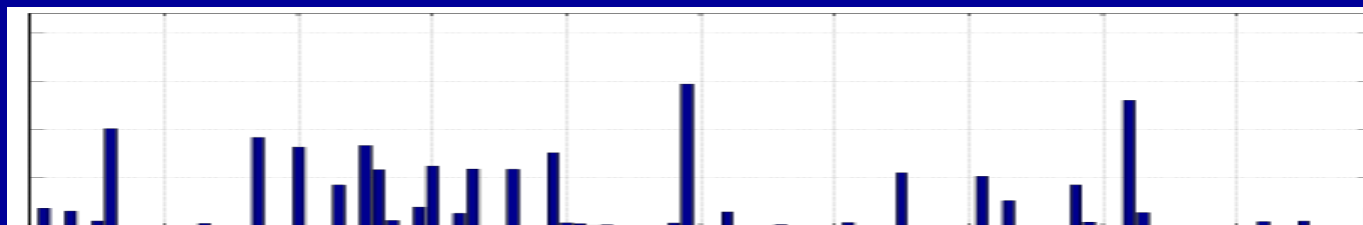
Rough Plastic



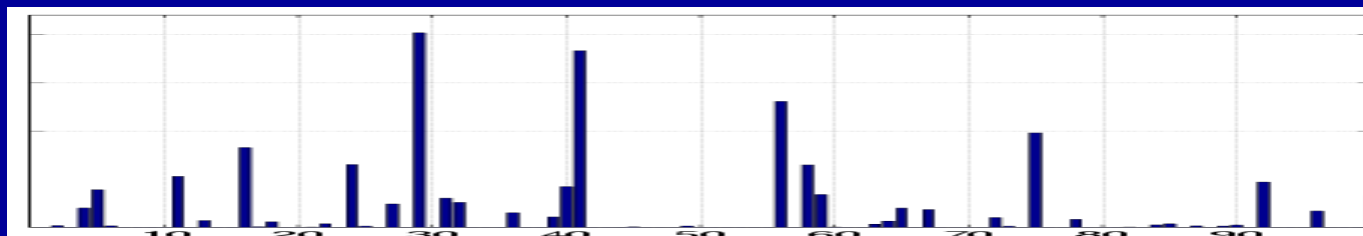
Pebbles



Plaster-b



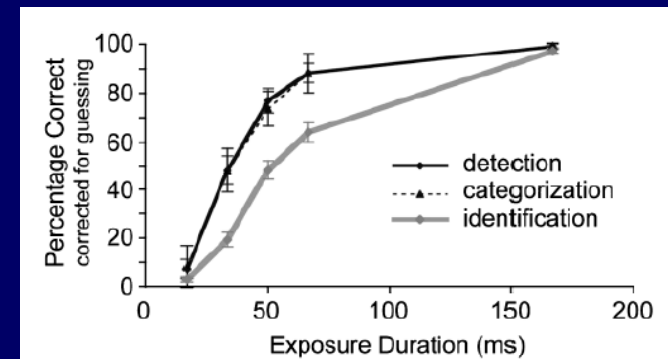
Terrycloth



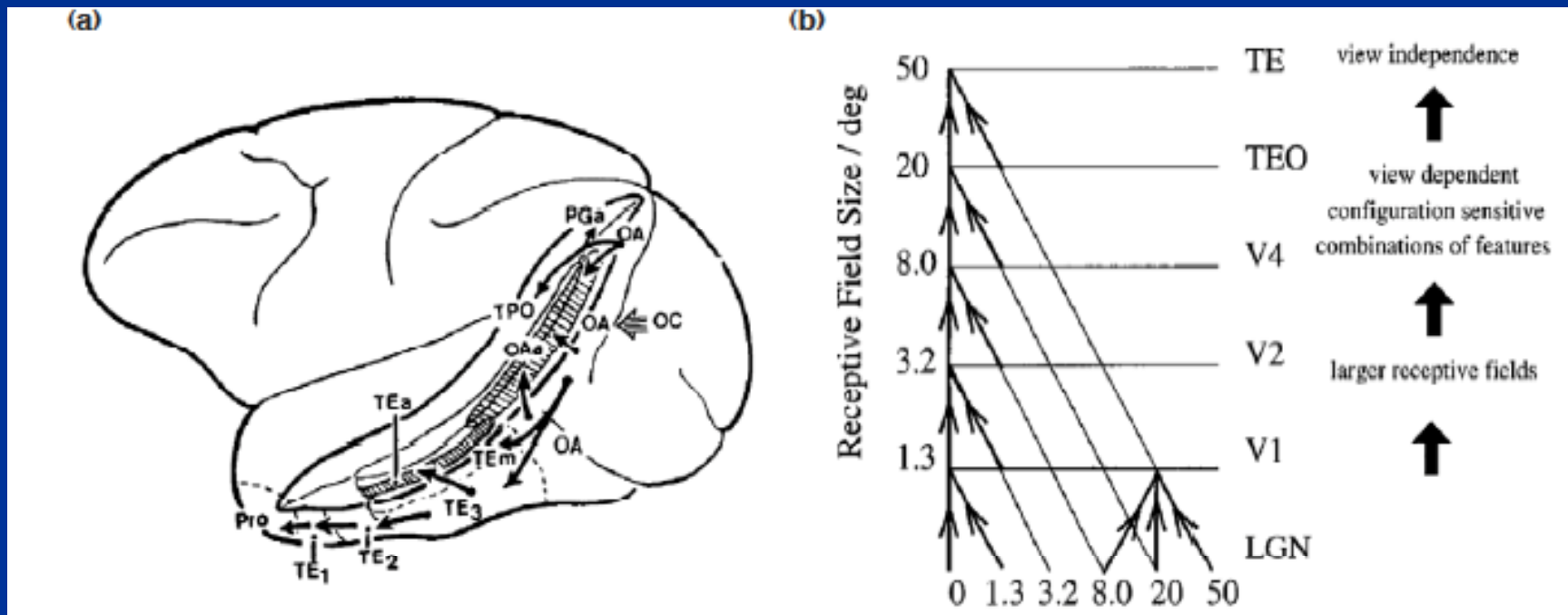


# Object Detection can be very fast

- On a task of judging animal vs no animal, humans can make mostly correct saccades in 150 ms (Kirchner & Thorpe, 2006)
  - Comparable to synaptic delay in the retina, LGN, V1, V2, V4, IT pathway.
  - Doesn't rule out feed back but shows **feed forward only is very powerful**
- Detection and categorization are practically simultaneous (Grill-Spector & Kanwisher, 2005)



# Rolls et al (2000)



# 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

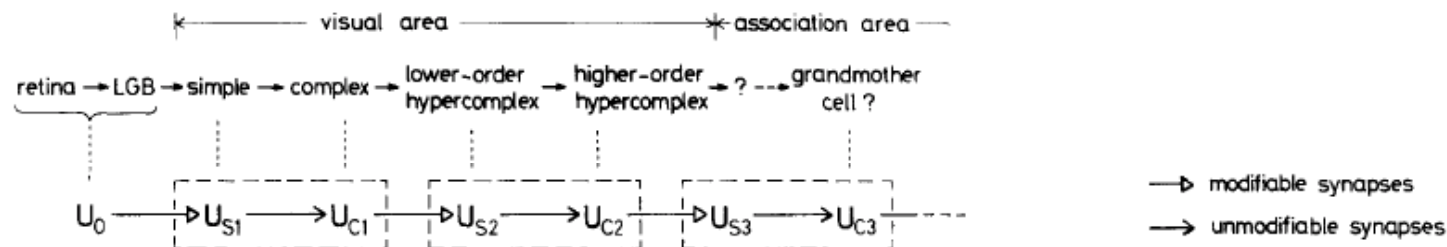


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

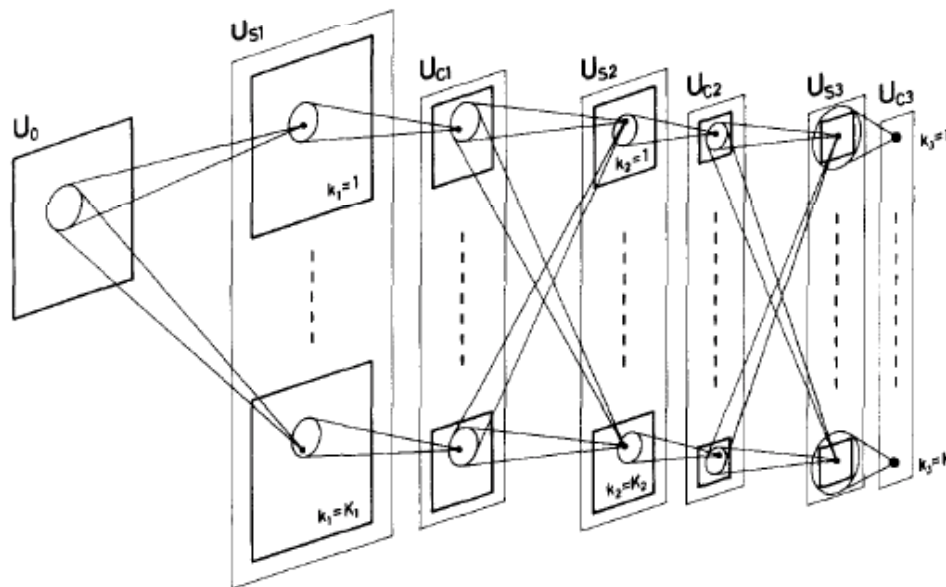
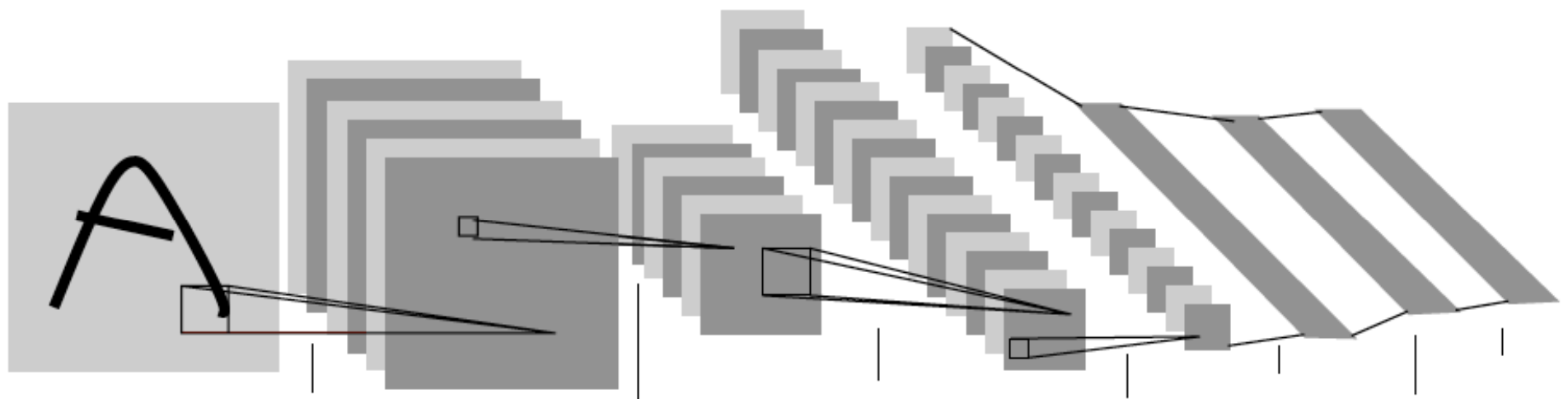


Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Biol. Cybernetics 36, 193–202 (1980)

# Convolutional Neural Networks (LeCun et al)

---



# A brief history of computer vision ..

Those who cannot remember the past are condemned to repeat it  
-George Santayana

# Fifty years of computer vision 1963-2013

- 1960s: Beginnings in artificial intelligence, image processing and pattern recognition
- 1970s: Foundational work on image formation: Horn, Koenderink, Longuet-Higgins ...
- 1980s: Vision as applied mathematics: geometry, multi-scale analysis, probabilistic modeling, control theory, optimization
- 1990s: Geometric analysis largely completed, vision meets graphics, statistical learning approaches resurface
- 2000s: Significant advances in visual recognition, range of practical applications

# Object recognition in computer vision

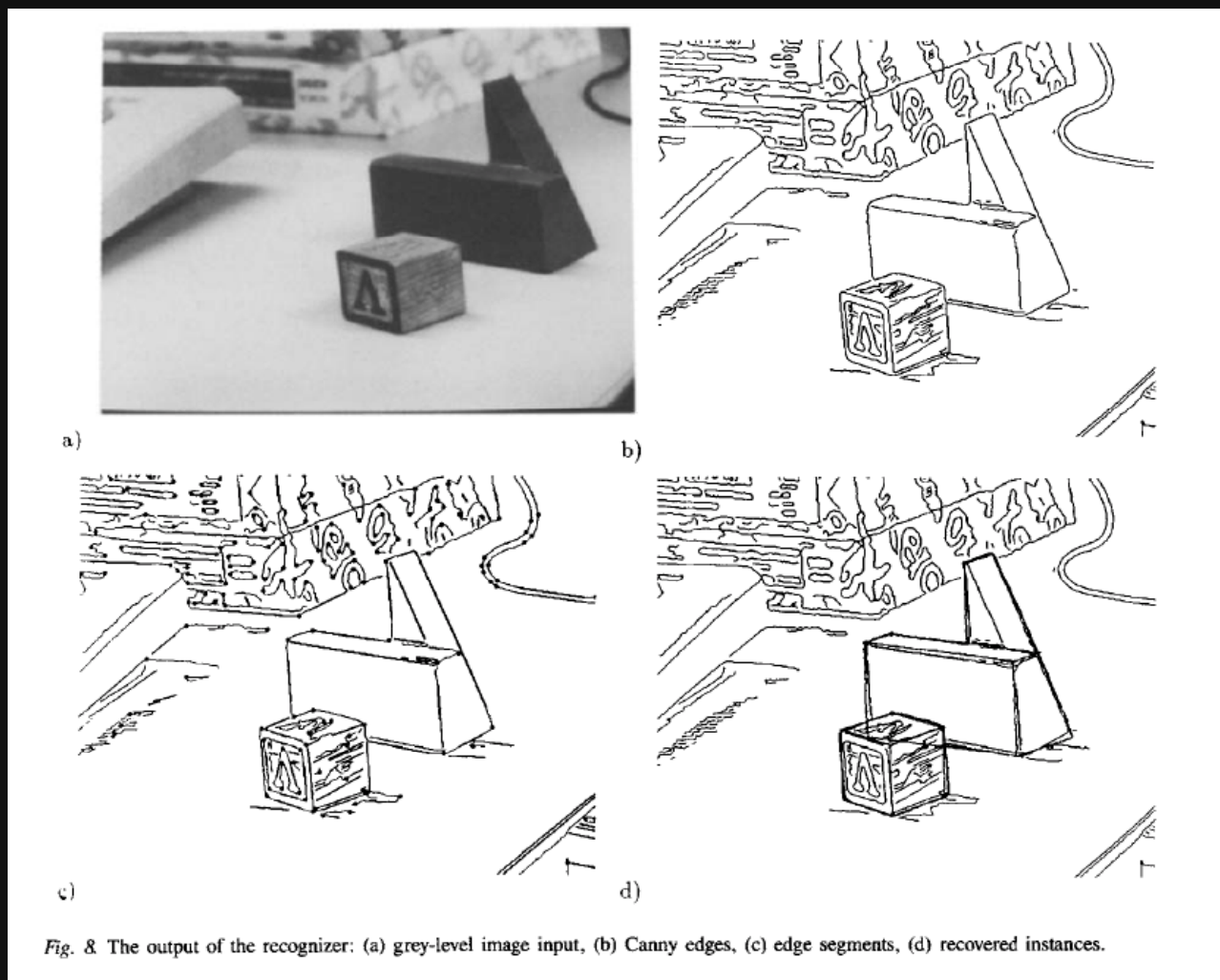
- Recognition as Pose Estimation
- Recognition as Description using Volumetric primitives
- Recognition as Pattern Classification
- Recognition as Deformable Matching

# Recognition as Pose Estimation: Object as a set of points in 3D

- Roberts (1963) , Faugeras & Hebert (1983), Huttenlocher & Ullman (1987)
- Variants
  - Geometric Hashing : Lamdan & Wolfson (1988)
  - Pose Clustering : Stockman (1987), Olson (1994)
  - Linear Combination of Views: Basri & Ullman (1991)



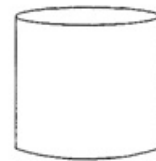
# Huttenlocher & Ullman's alignment Algorithm (1990)



# Recognition as Fitting Volumetric Primitives: Object as a hierarchy of simple shapes

- Binford (1971) , Marr & Nishihara (1978), Biederman(1987)
- Discredited as an approach for recognition in general, it has retained appeal for analyzing images of people

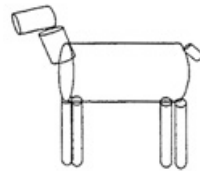
# The Stick Figure Ideal



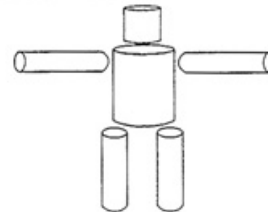
Zylinder



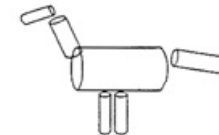
Körperglied



Vierbeiner



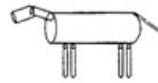
Zweibeiner



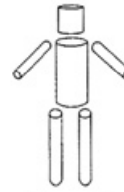
Vogel



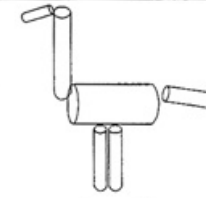
Dickes Glied



Hund



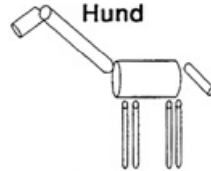
Mensch



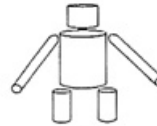
Strauß



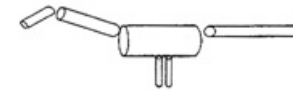
Dünnes Glied



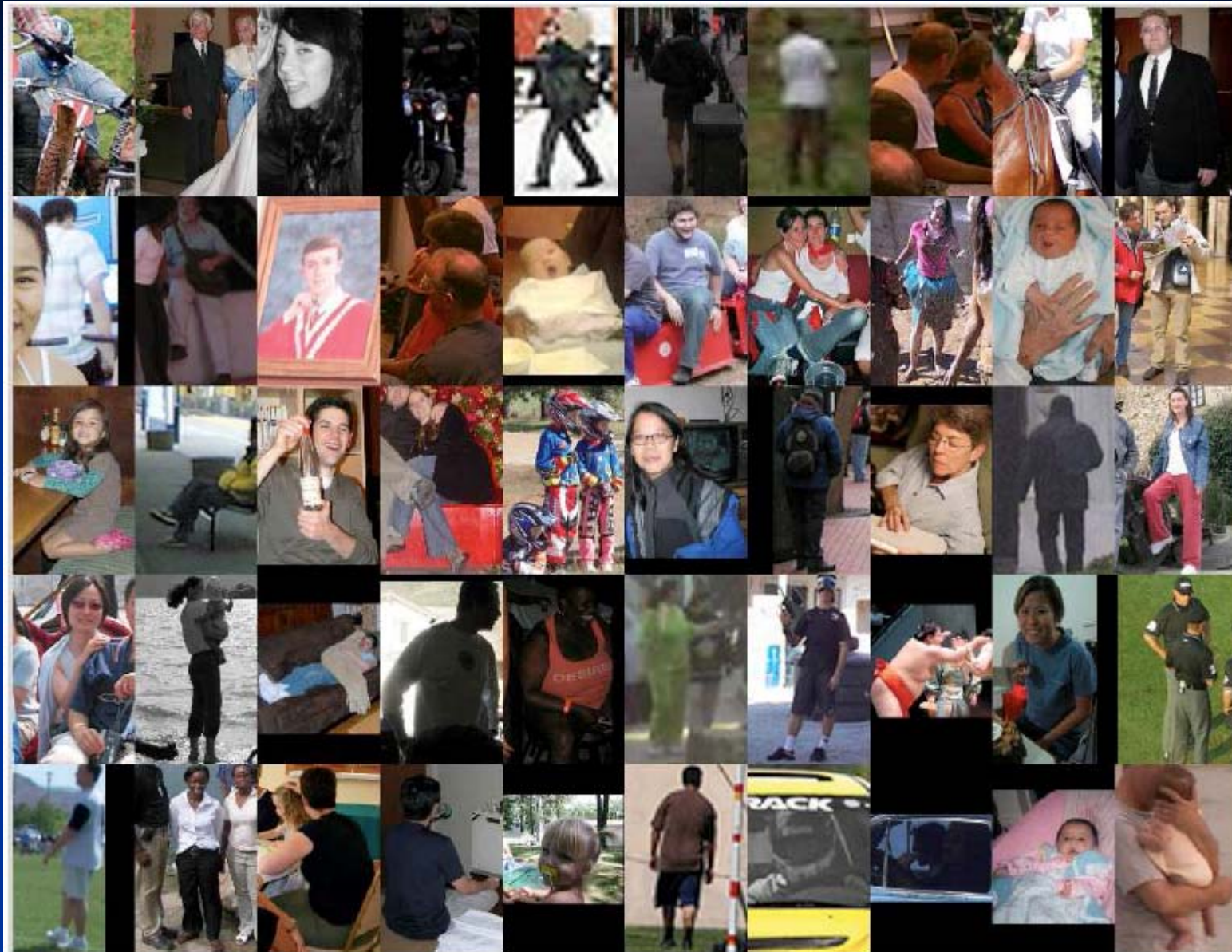
Giraffe



Affe



Taube



# Recognition as Statistical Pattern Classification: Object as a feature vector

- Optical Character Recognition studied as far back as the 1950s. Recent years focus on handwritten digit classification and face detection.
- Some examples:
  - Neural networks: Neocognitron (Fukushima, 1980, 1988) , Convolution Neural Networks (LeCun et al), C2 Features (Serre, Wolf & Poggio 2005)
  - Support Vector Machines (various)
  - Decision Trees (Amit, Geman, & Wilder, 1997)
  - Boosted Decision Trees (Viola & Jones, 2001)

3 6 8 1 7 9 6 6 9 1  
6 7 5 7 8 6 3 4 8 5  
2 1 7 9 7 1 2 8 4 5  
4 8 1 9 0 1 8 8 9 4  
7 6 1 8 6 4 1 5 6 0  
7 5 9 2 6 5 8 1 9 7  
2 2 2 2 2 3 4 4 8 0  
0 2 3 8 0 7 3 8 5 7  
0 1 4 6 4 6 0 2 4 3  
7 1 2 8 7 6 9 8 6 1

Fig. 4. Size-normalized examples from the MNIST database.

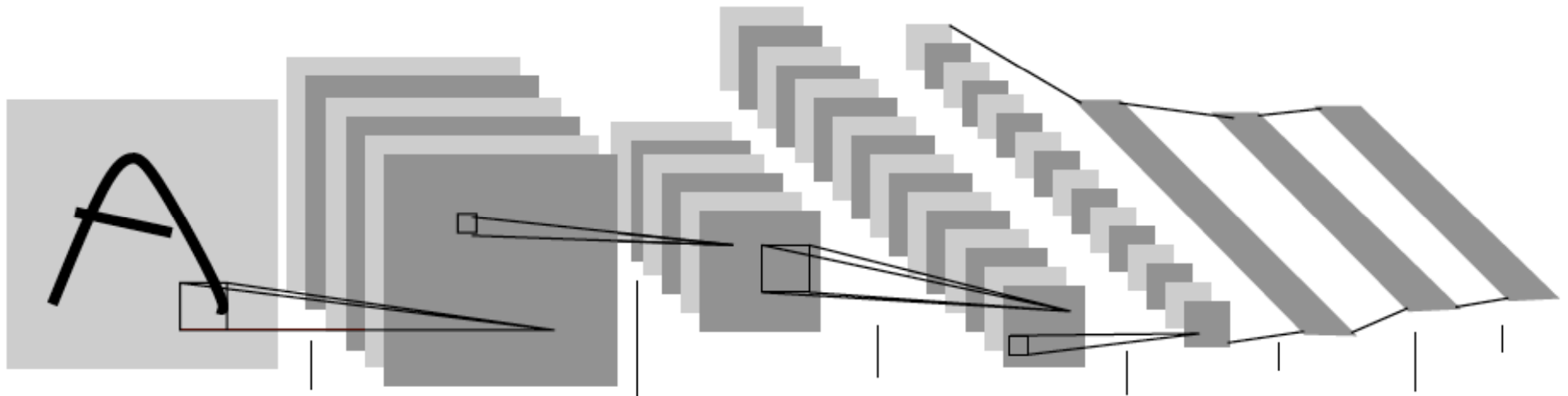
# Handwritten digit recognition (MNIST,USPS)



- LeCun's Convolutional Neural Networks variations (0.8%, 0.6% and 0.4% on MNIST)
- Tangent Distance (Simard, LeCun & Denker: 2.5% on USPS)
- Randomized Decision Trees (Amit, Geman & Wilder, 0.8%)
- K-NN based Shape context/TPS matching (Belongie, Malik & Puzicha: 0.6% on MNIST)

# Convolutional Neural Networks (LeCun et al)

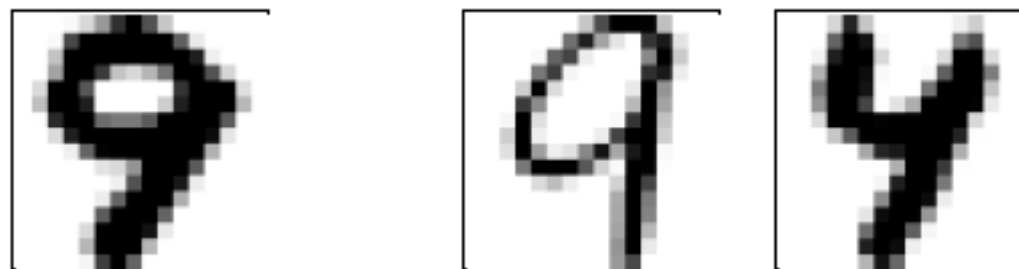
---



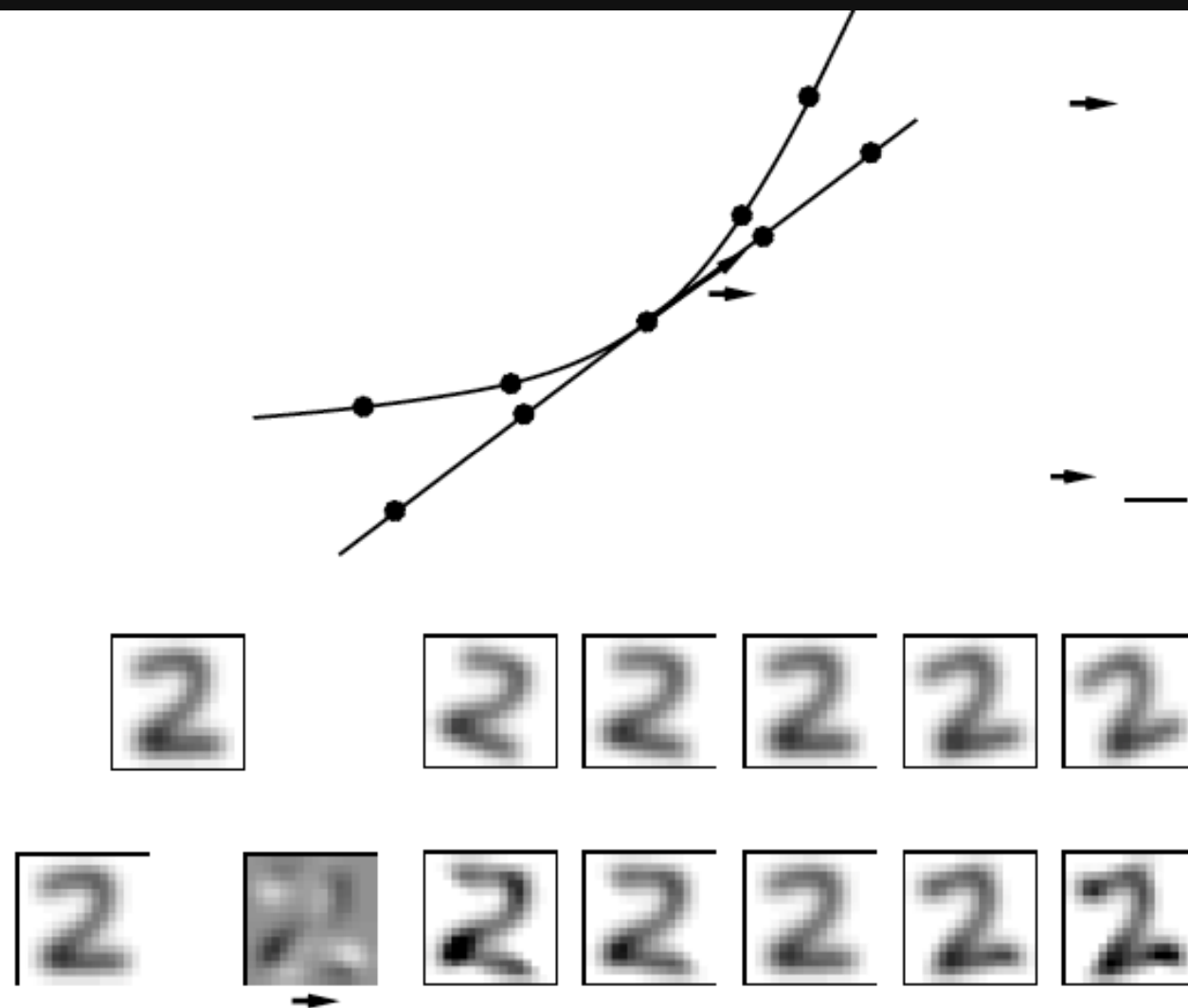


# The idea behind Tangent Distance (Simard et al)

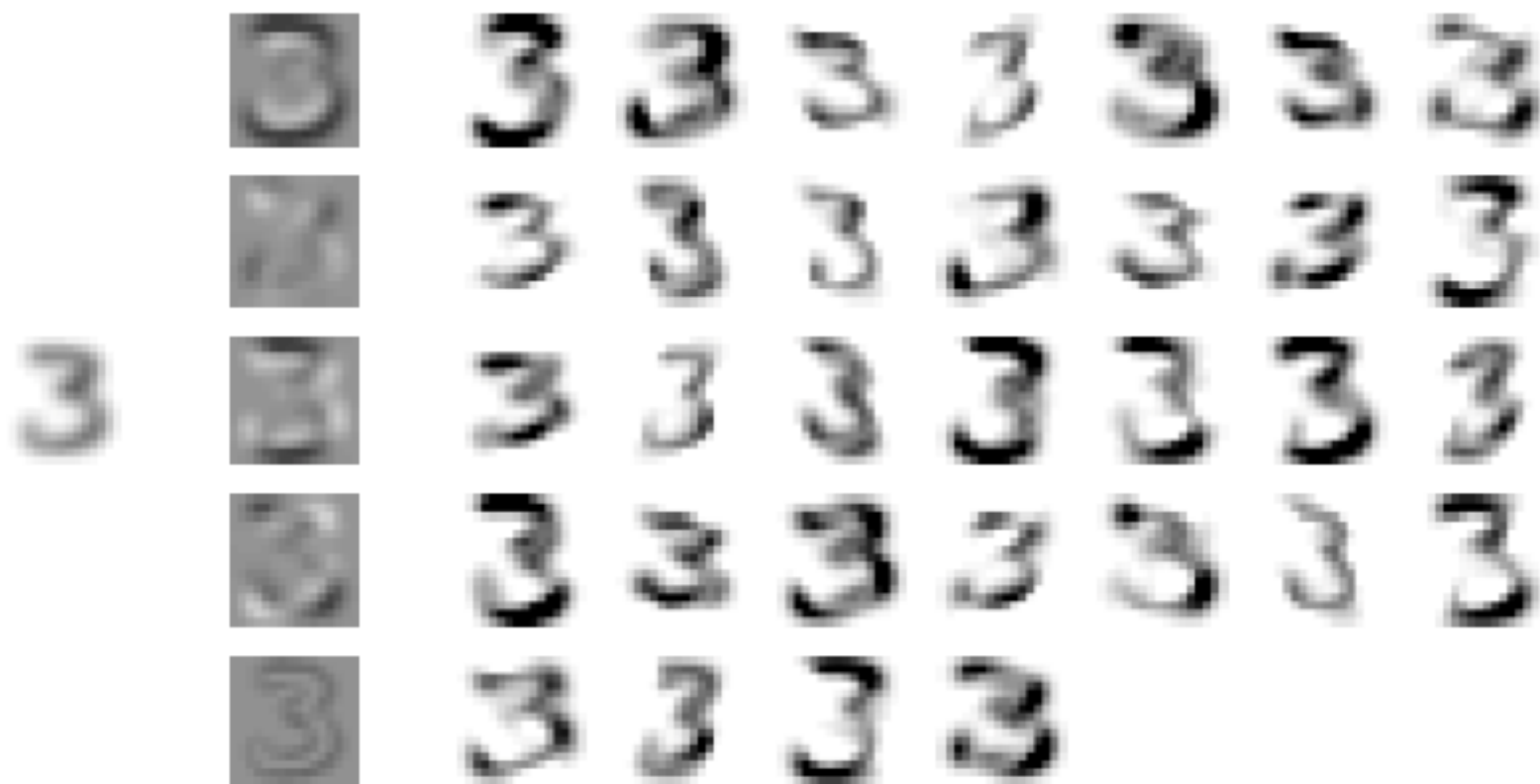
---



**Fig. 1.** According to the Euclidean distance the pattern to be classified is more similar to prototype B. A better distance measure would find that prototype A is closer because it differs mainly by a rotation and a thickness transformation, two transformations which should leave the classification invariant.



**Fig. 2.** Top: Representation of the effect of the rotation in pixel space. Middle: Small rotations of an original digitized image of the digit “2”, for different angle values of  $\alpha$ . Bottom: Images obtained by moving along the tangent to the transformation curve for the same original digitized image  $P$  by adding various amounts ( $\alpha$ ) of the tangent vector  $T$ .



**Fig. 6.** Left: Original image. Middle: 5 tangent vectors corresponding respectively to the 5 transformations: scaling, rotation, expansion of the X axis while compressing the Y axis, expansion of the first diagonal while compressing the second diagonal and thickening. Right: 32 points in the tangent space generated by adding or subtracting each of the 5 tangent vectors.

# Amit, Geman & Wilder (1997)

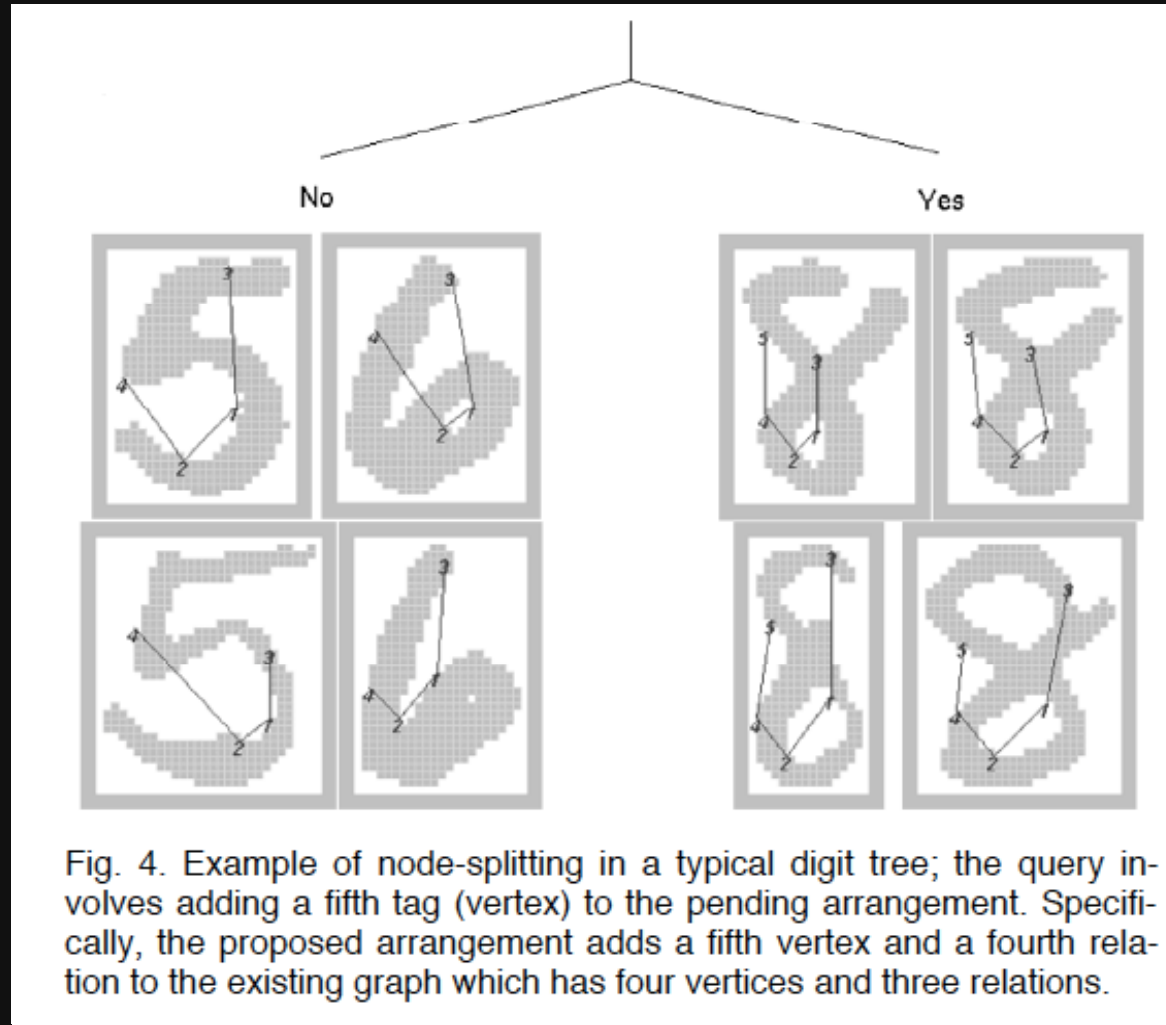
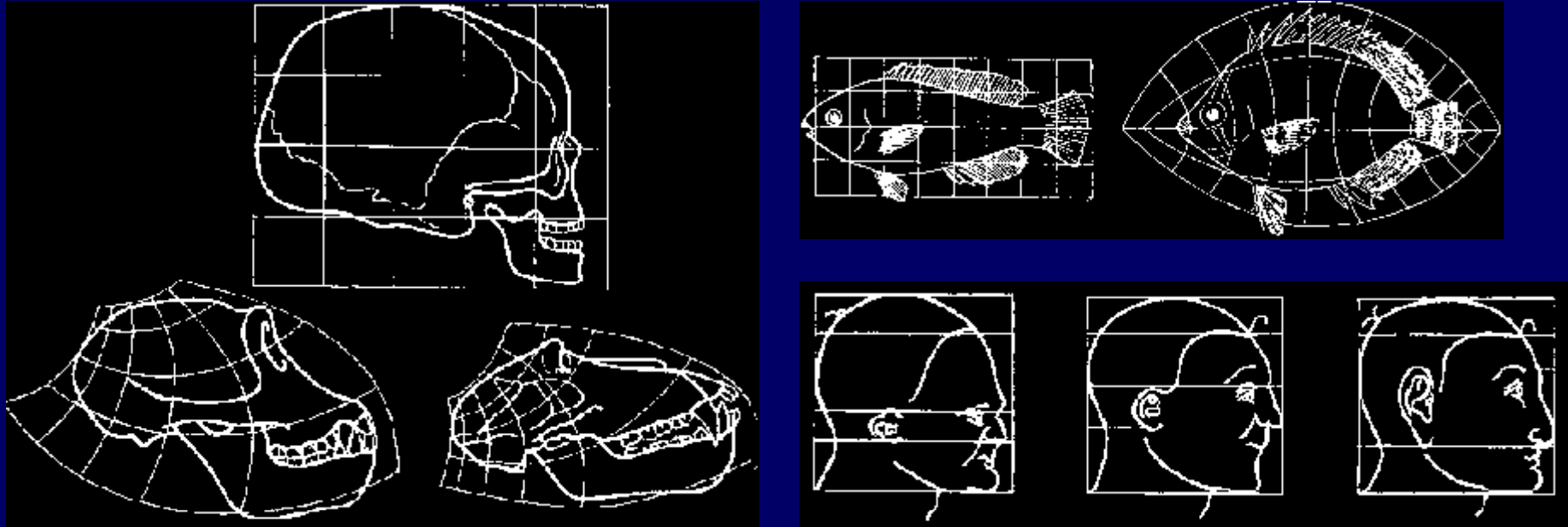


Fig. 4. Example of node-splitting in a typical digit tree; the query involves adding a fifth tag (vertex) to the pending arrangement. Specifically, the proposed arrangement adds a fifth vertex and a fourth relation to the existing graph which has four vertices and three relations.

# Recognition as Pictorial Structure Matching: Object as a spatial configuration of features

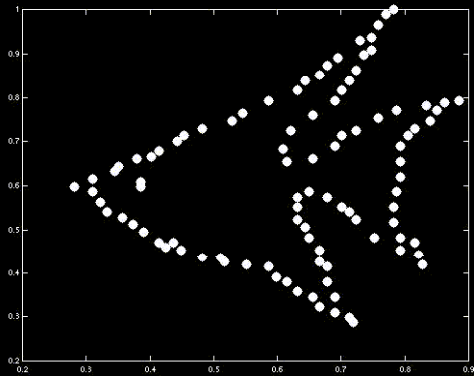
- Transformations to model shape variation - D'Arcy Wentworth Thompson (1910)
- Grenander (1970s and later) probabilistic models on transformations
- Fischler and Elschlager (1973) - deformable matching of landmarks, “point masses”, in a configuration of “springs” to model deformable templates.
- Von der Malsburg - dynamic link architecture for neural modeling, elastic graph matching for face recognition (1993, 1997)
- Felzenszwalb and Huttenlocher (2000) - pictorial structures for aligning human bodies to stick figures using dynamic programming
- Belongie, Malik & Puzicha (2001) use “shape contexts” as point descriptors, and thin plate splines to model deformation.

# Modeling shape variation in a category

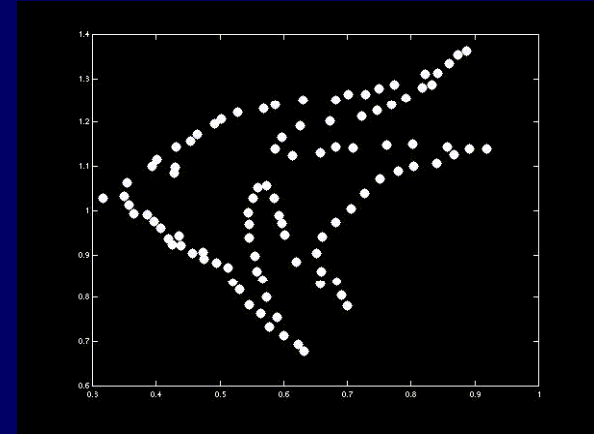


- D'Arcy Thompson: *On Growth and Form*, 1917
  - studied transformations between shapes of organisms

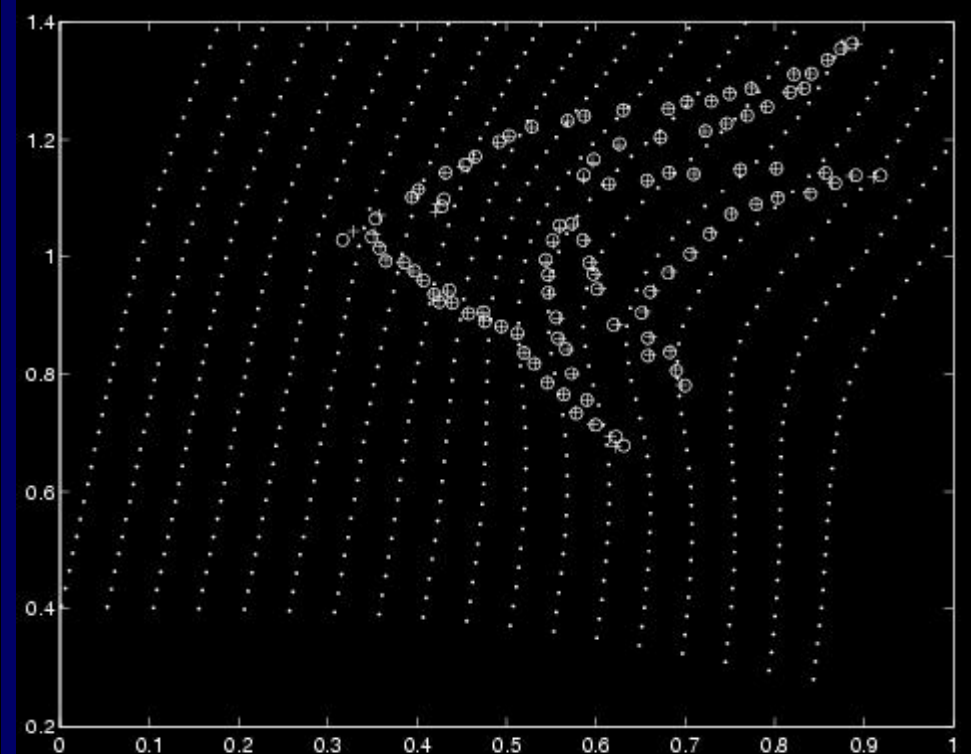
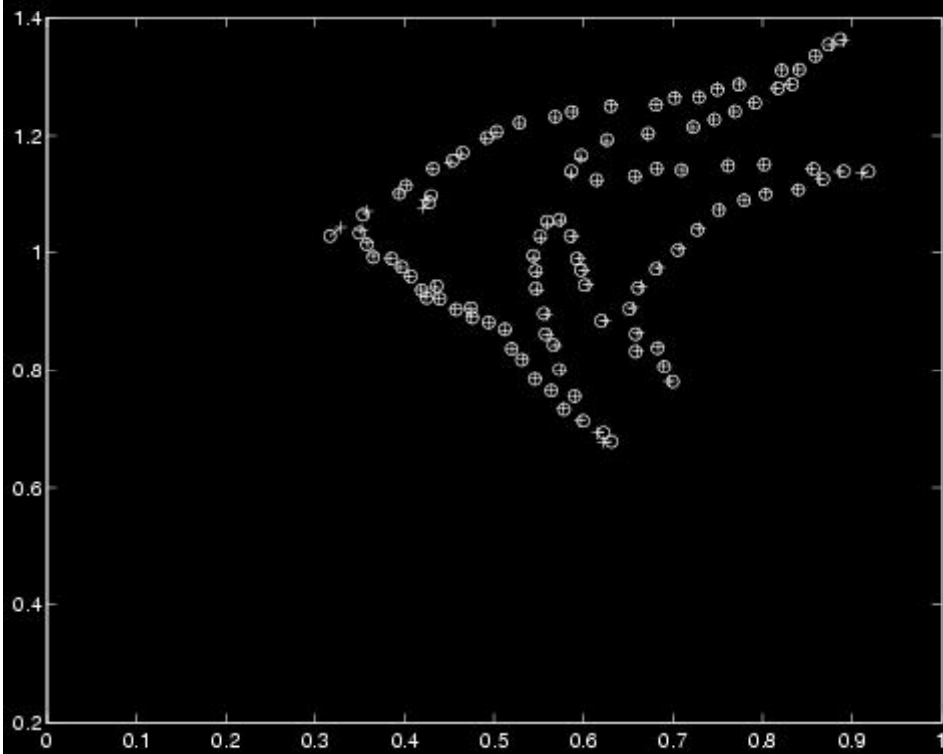
# Matching Example



model



target



# EZ-Gimpy Results (Mori & Malik, 2003)

- 171 of 192 images correctly identified: 92 %



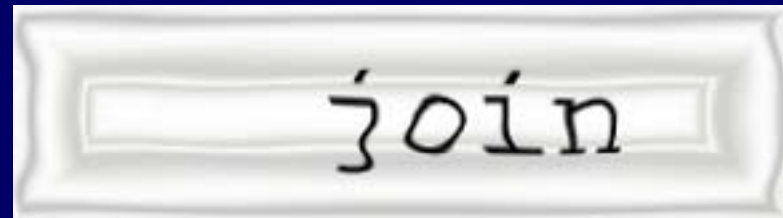
horse



spade



smile



join



canvas

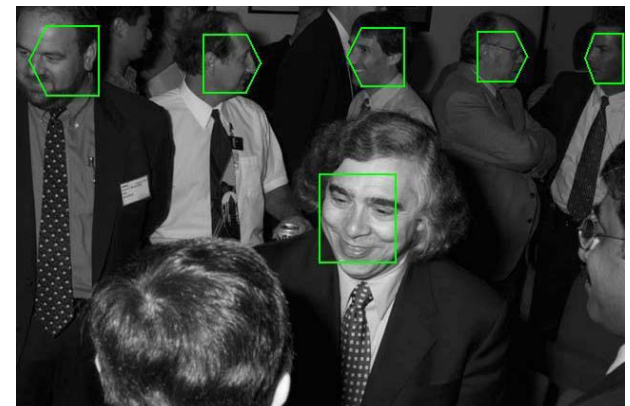
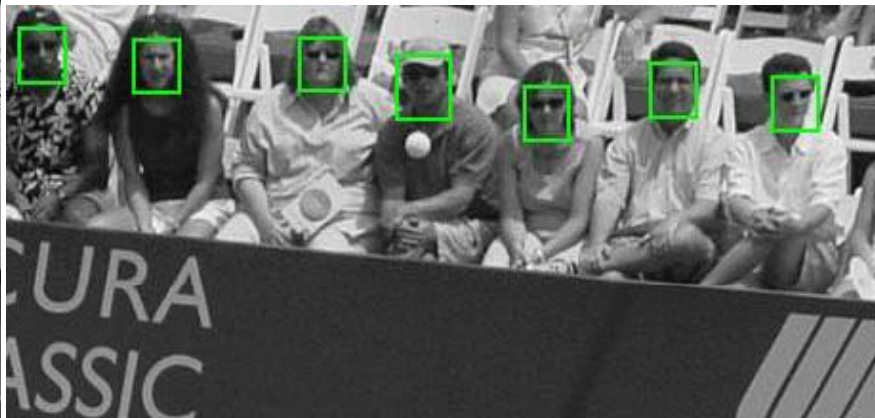
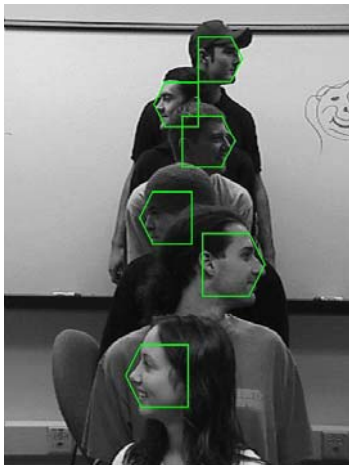
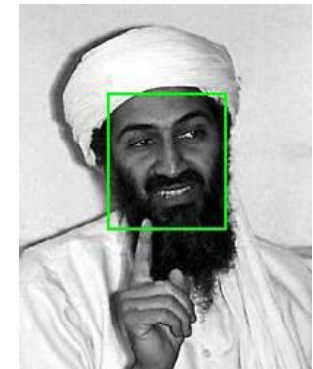
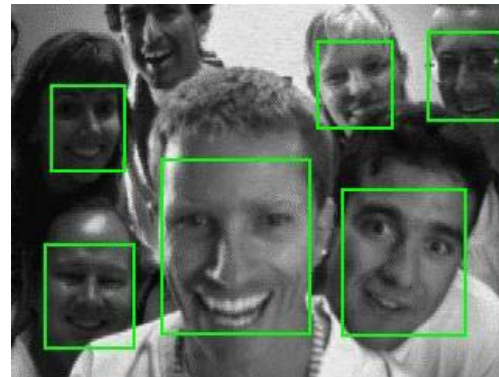
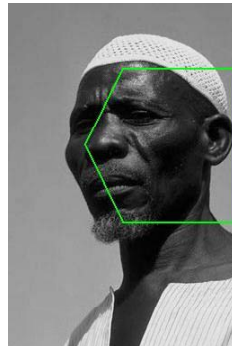
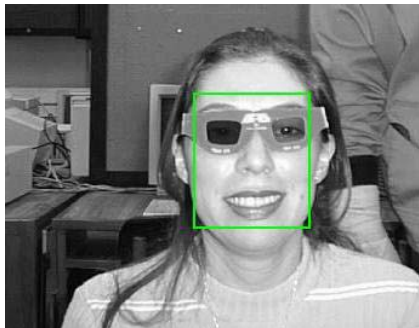


here

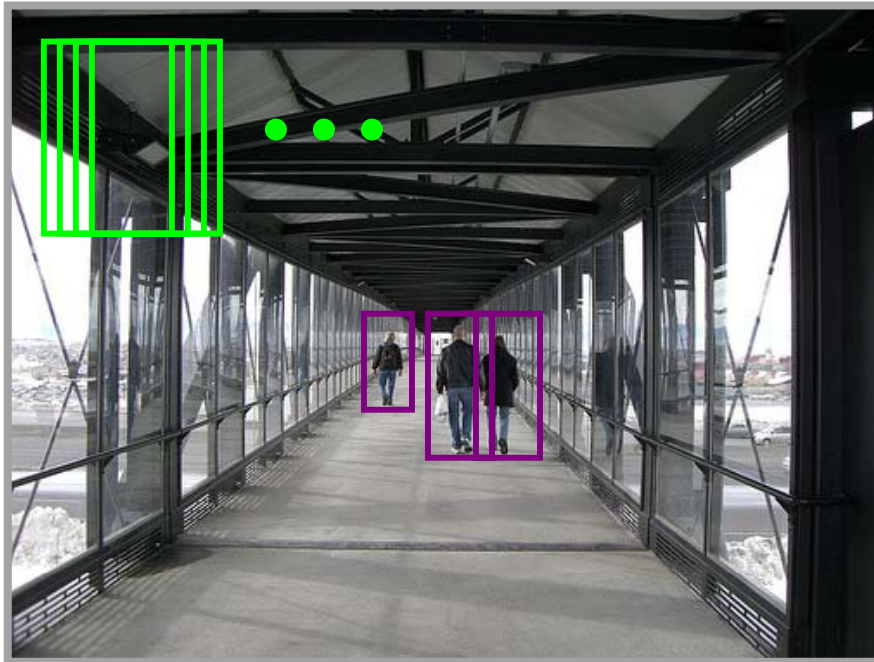


# Face Detection

Carnegie Mellon University



# Multiscale sliding window



Paradigm introduced by Rowley, Baluja & Kanade 96 for face detection  
Viola & Jones 01, Dalal & Triggs 05, Felzenszwalb, McAllester, Ramanan 08





# The PASCAL Visual Object Classes Challenge 2010 (VOC2010)

## Part 2 – Detection Task

Mark Everingham

Luc Van Gool

Chris Williams

John Winn

Andrew Zisserman

# PASCAL Visual Object Challenge

## Dining Table



## Dog



## Horse



## Motorbike



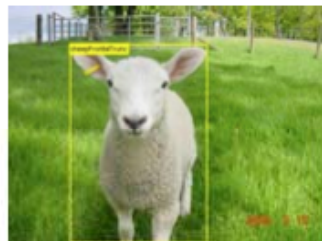
## Person



## Potted Plant



## Sheep



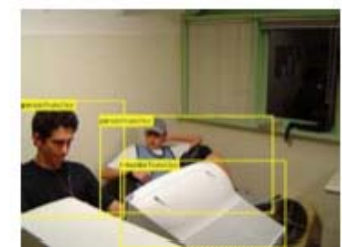
## Sofa



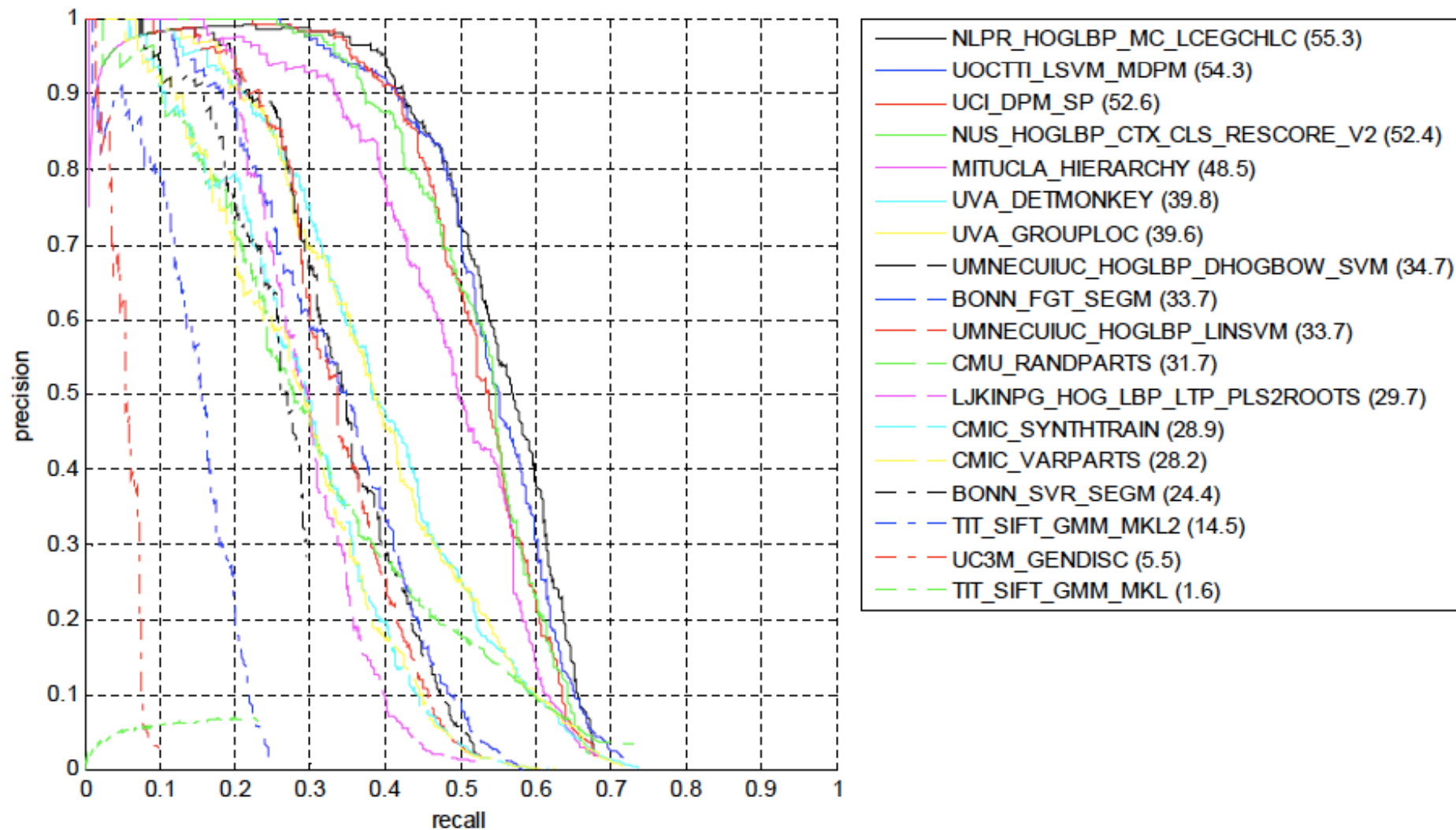
## Train



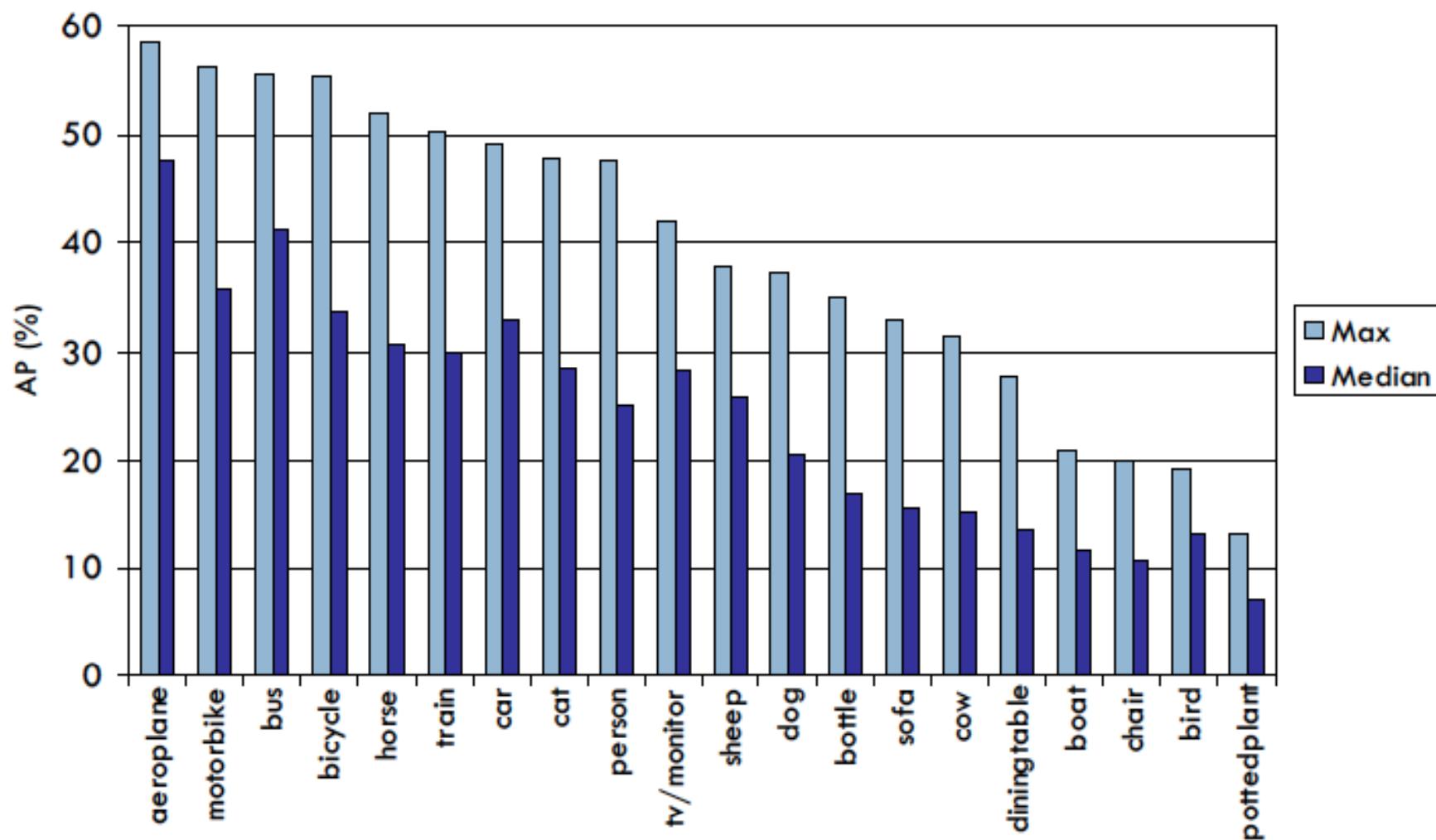
## TV/Monitor



# Precision/Recall - Bicycle



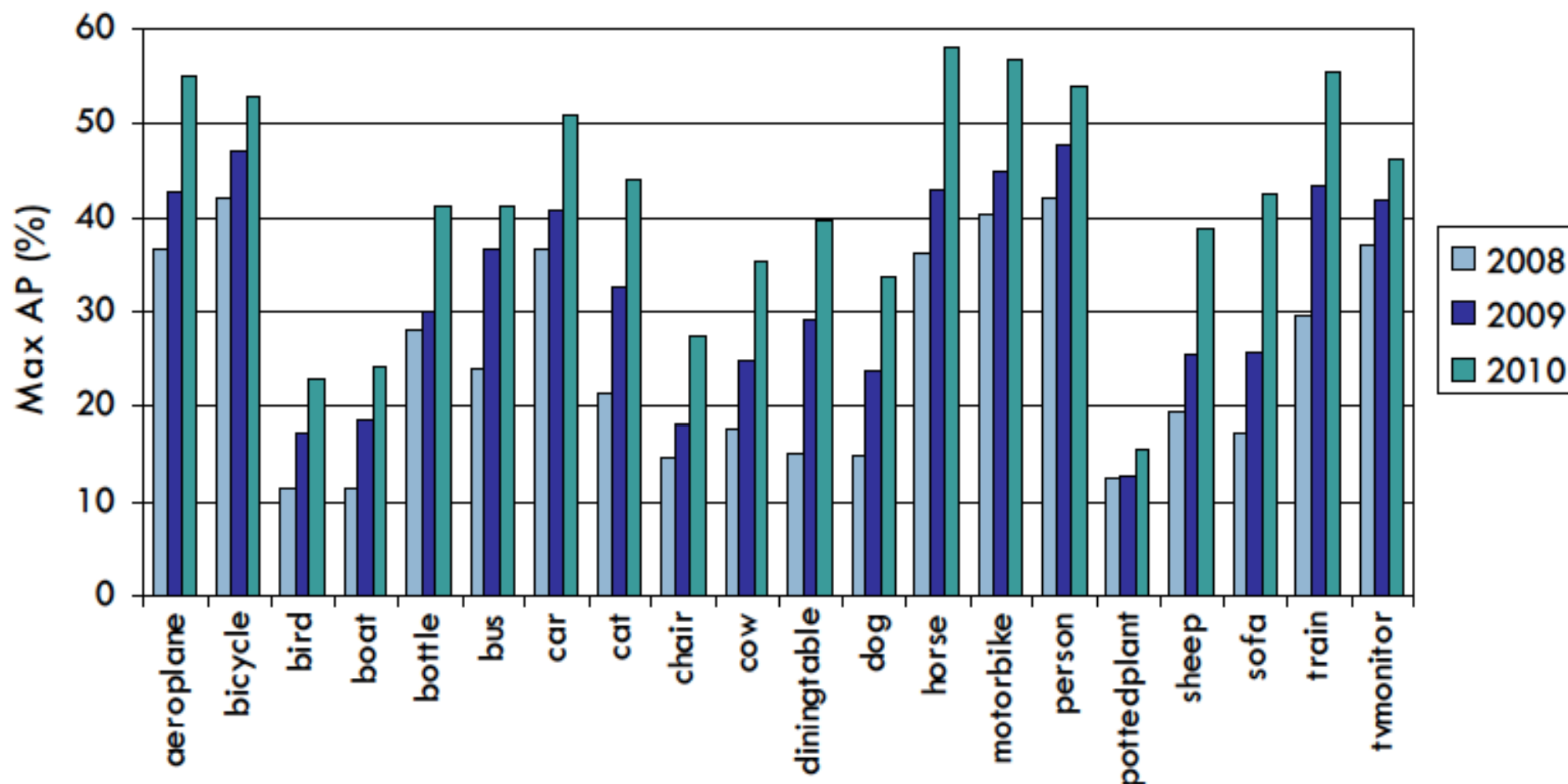
# AP by Class



- Max AP: 58.4% (aeroplane) ... 13.0% (potted plant)

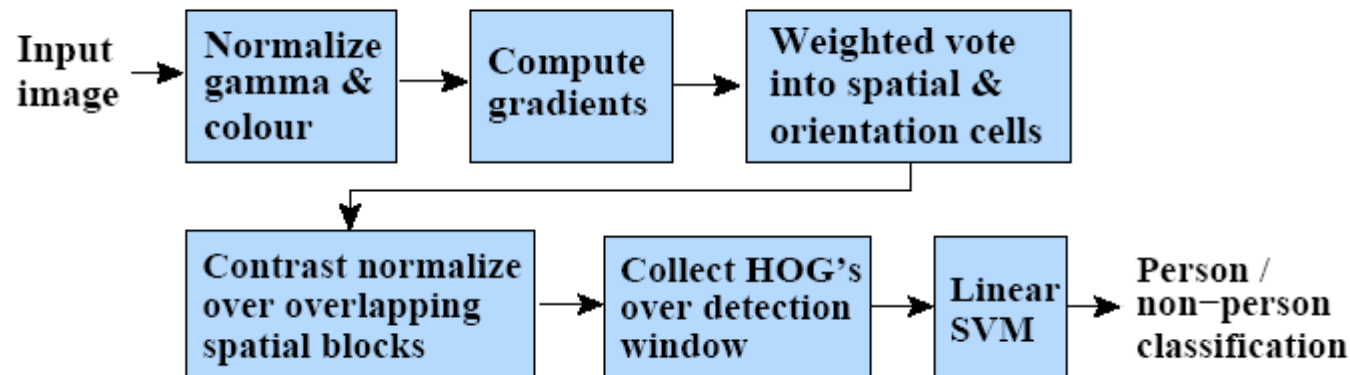
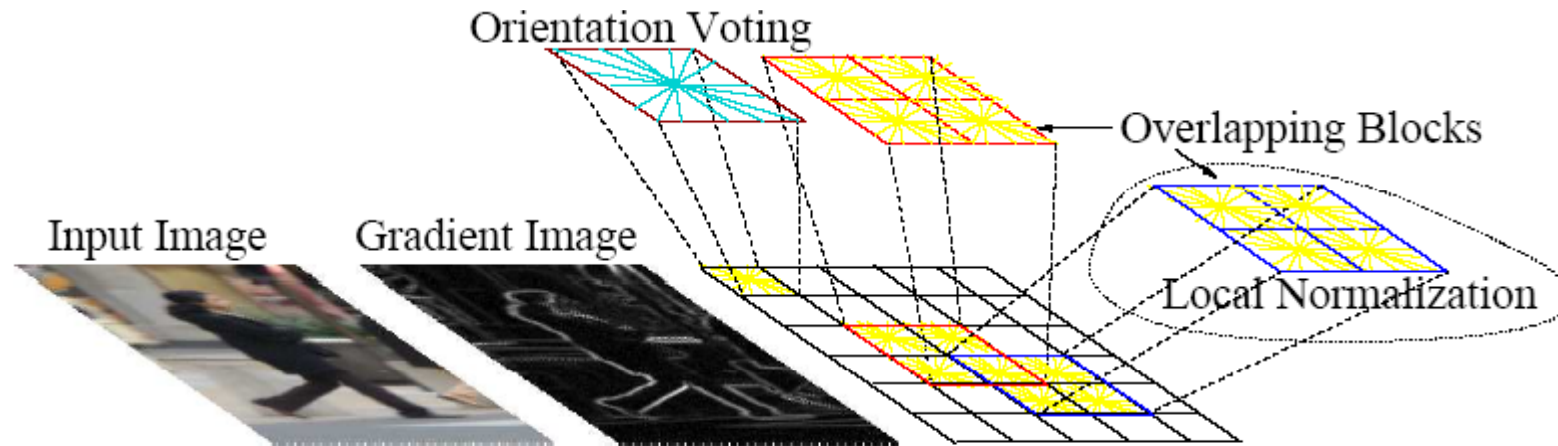


# Progress 2008-2010



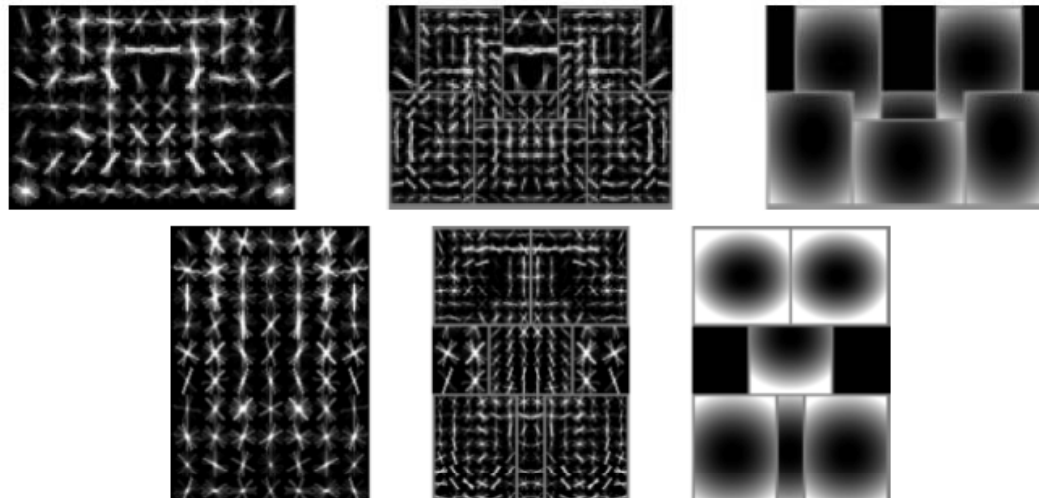
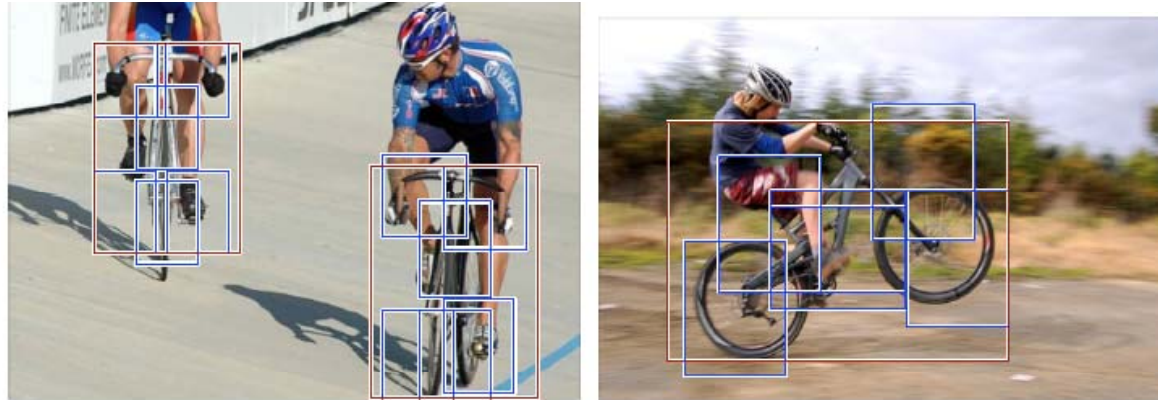
- Results on 2008 data improve for best 2009 and 2010 methods for all categories, by over 100% for some categories
  - Caveat: Better methods or more training data?

# A good building block is a linear SVM trained on HOG features (Dalal & Triggs)



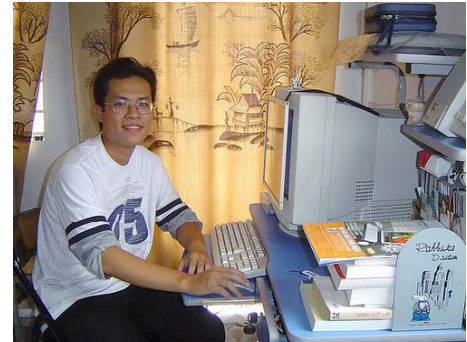
# Object Detection with Discriminatively Trained Part Based Models

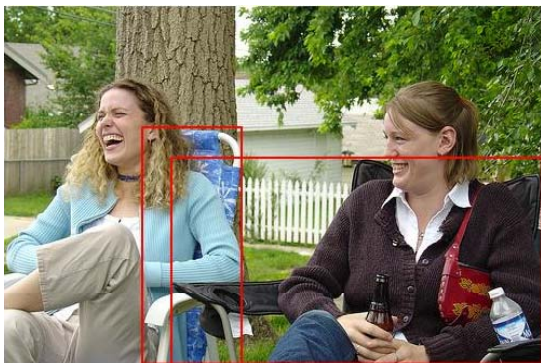
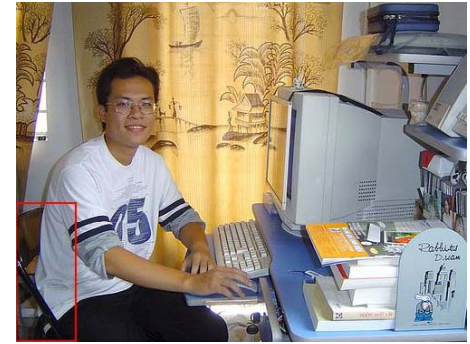
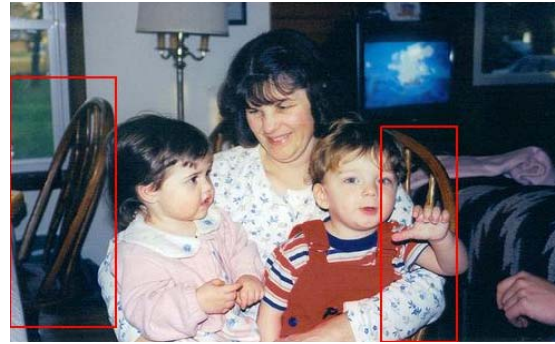
Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan





AP=0.23





# Datasets and computer vision

## (slide credit: Fei-Fei Li)



**UIUC Cars (2004)**  
S. Agarwal, A. Awan, D. Roth



**CMU/VASC Faces (1998)**  
H. Rowley, S. Baluja, T. Kanade



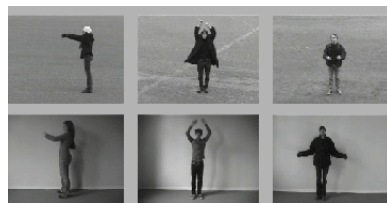
**FERET Faces (1998)**  
P. Phillips, H. Wechsler, J. Huang, P. Raus



**COIL Objects (1996)**  
S. Nene, S. Nayar, H. Murase



**MNIST digits (1998-10)**  
Y LeCun & C. Cortes



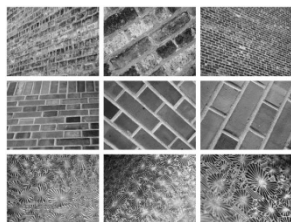
**KTH human action (2004)**  
I. Leptev & B. Caputo



**Sign Language (2008)**  
P. Buehler, M. Everingham, A. Zisserman



**Segmentation (2001)**  
D. Martin, C. Fowlkes, D. Tal, J. Malik.



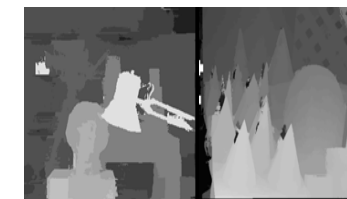
**3D Textures (2005)**  
S. Lazebnik, C. Schmid, J. Ponce



**CuRET Textures (1999)**  
K. Dana B. Van Ginneken S. Nayar J. Koenderink



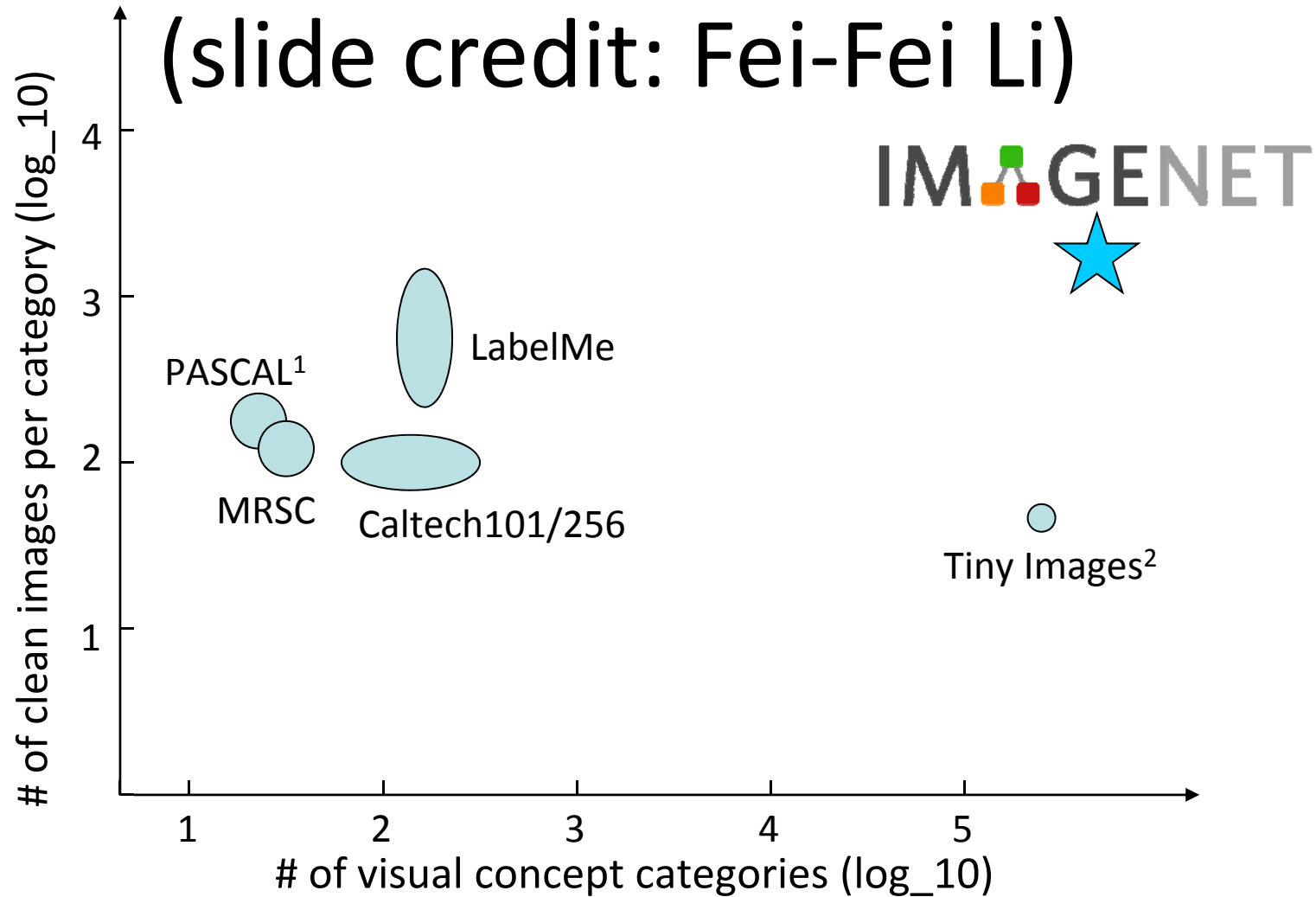
**CAVIAR Tracking (2005)**  
R. Fisher, J. Santos-Victor J. Crowley



**Middlebury Stereo (2002)**  
D. Scharstein R. Szeliski

# Comparison among free datasets


(slide credit: Fei-Fei Li)



1. Excluding the Caltech101 datasets from PASCAL
2. No image in this dataset is human annotated. The # of clean images per category is a rough estimation



# The more you look, the more you see!

Image shown to subjects	40ms	80ms	107ms	500ms
	<p>“Possibly outdoor scene, maybe a farm. I could not tell for sure.”</p>	<p>“There seem to be two people in the center of the scene.”</p>	<p>“ People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight.”</p>	<p>“Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets...”</p>
<p>Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].</p>				

# So much remains to be done...

- Objects, Scenes, Events
- The semantic gap is to be confronted, not avoided!