What is a Chair?

The object

The texture

The object

The texture

The object

The scene

Instances vs. categories

Instances Find these two toys

Can nail it

Categories Find a bottle:

Can’t do
unless you do not
care about few errors…

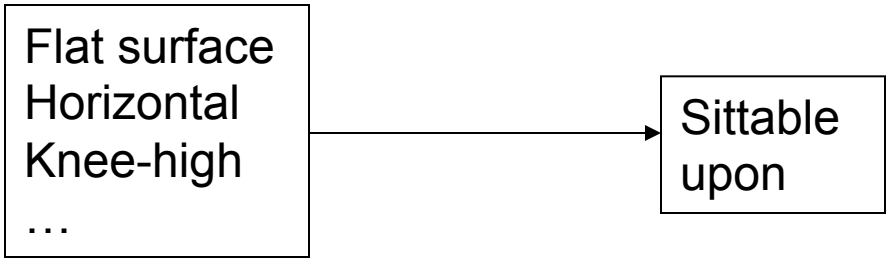
Why do we care about recognition?

Perception of function: We can perceive the 3D shape, texture, material properties, without knowing about objects. But, the concept of category encapsulates also information about what can we do with those objects.

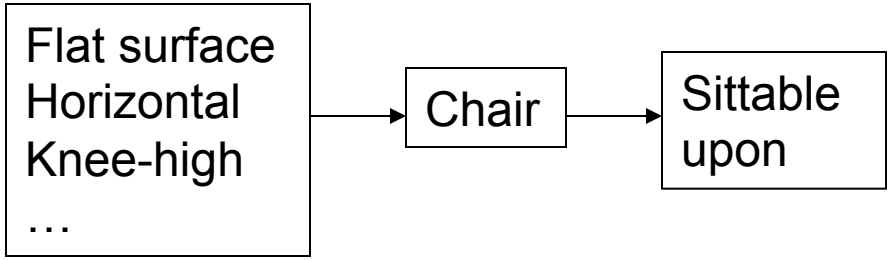
"We therefore include the perception of function as a proper –indeed, crucial- subject for vision science", *from Vision Science, chapter 9, Palmer*.

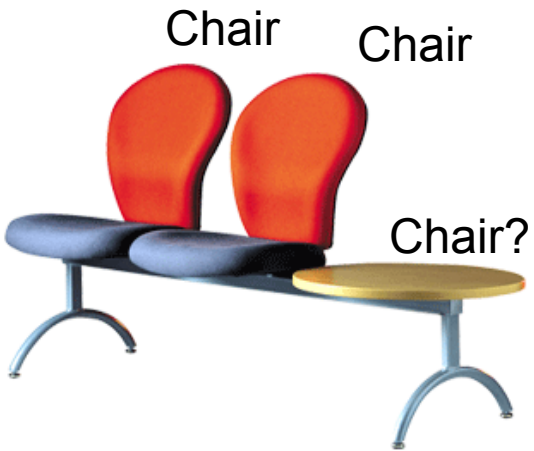
The perception of function

- Direct perception (affordances): Gibson



- Mediated perception (Categorization)

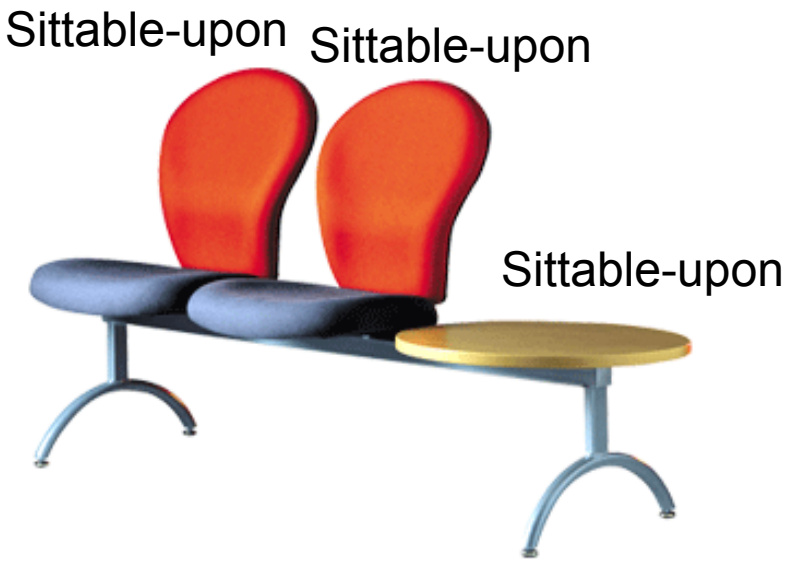




Direct perception

Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)



It does not seem easy to sit-upon this…

Direct perception

Some aspects of an object function can be perceived directly

- Observer relativity: Function is observer dependent

Limitations of Direct Perception

Objects of similar structure might have very different functions

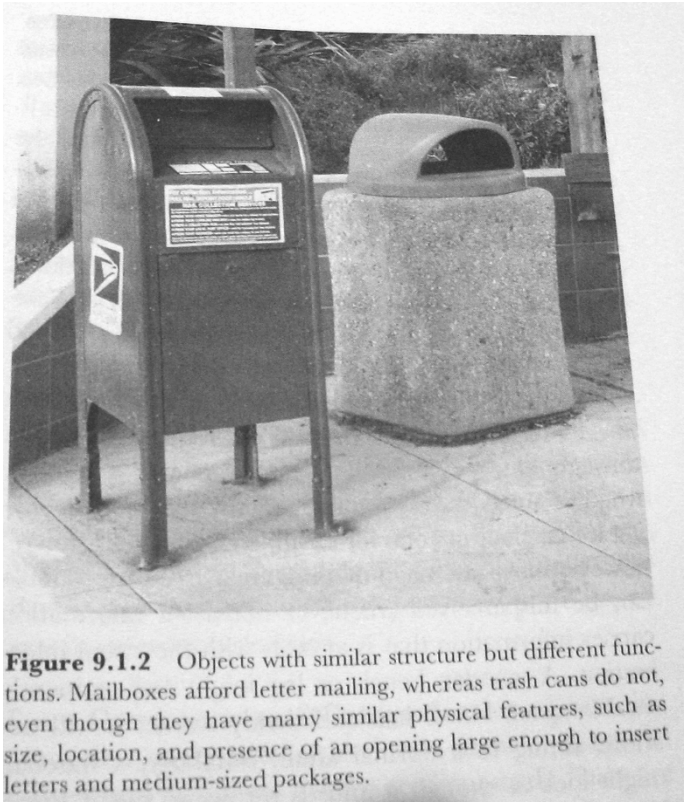


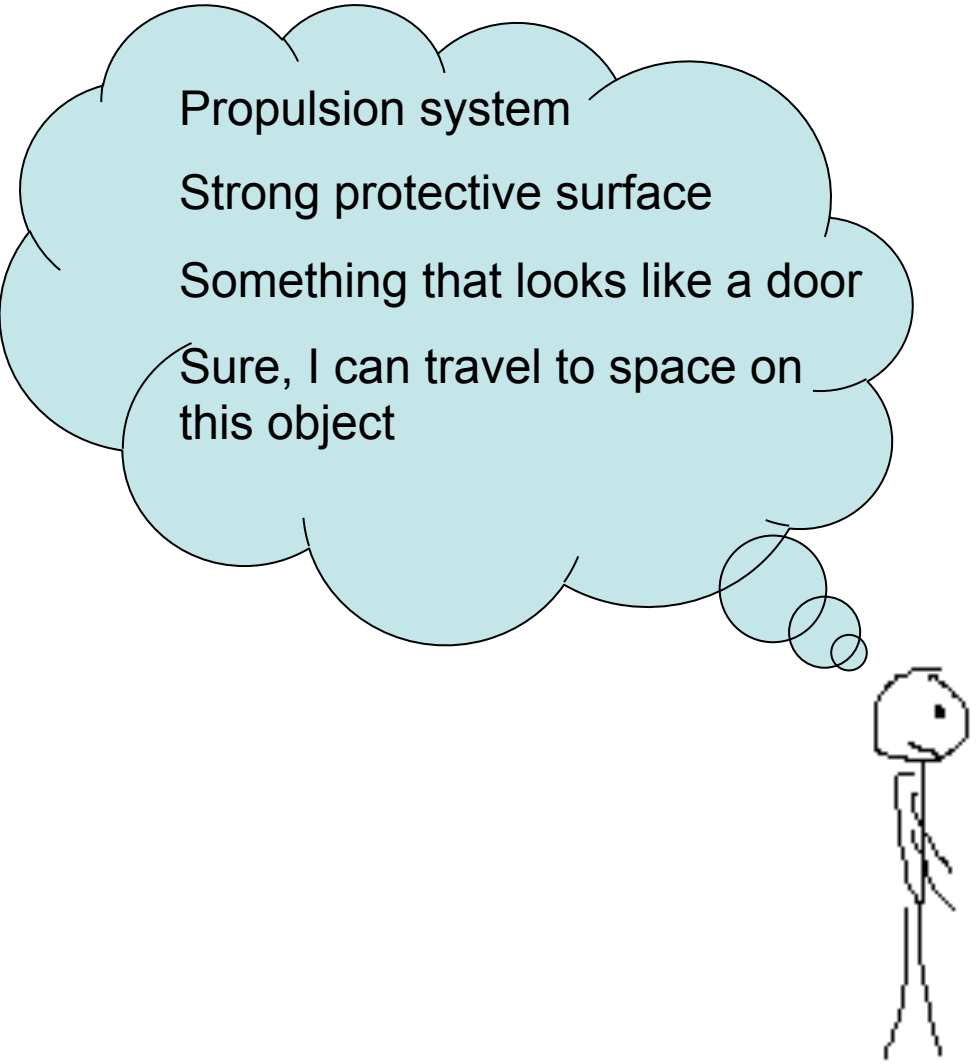
Figure 9.1.2 Objects with similar structure but different functions. Mailboxes afford letter mailing, whereas trash cans do not, even though they have many similar physical features, such as size, location, and presence of an opening large enough to insert letters and medium-sized packages.

Not all functions seem to be available from direct visual information only.

The functions are the same at some level of description: we can put things inside in both and somebody will come later to empty them. However, we are not expected to put inside the same kinds of things…

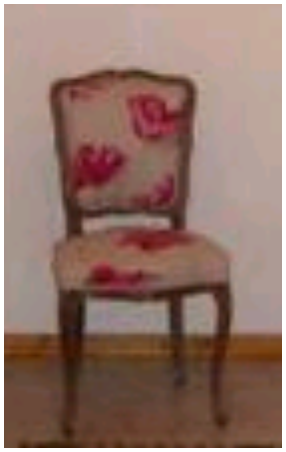
Limitations of Direct Perception

Visual appearance might be a very weak cue to function

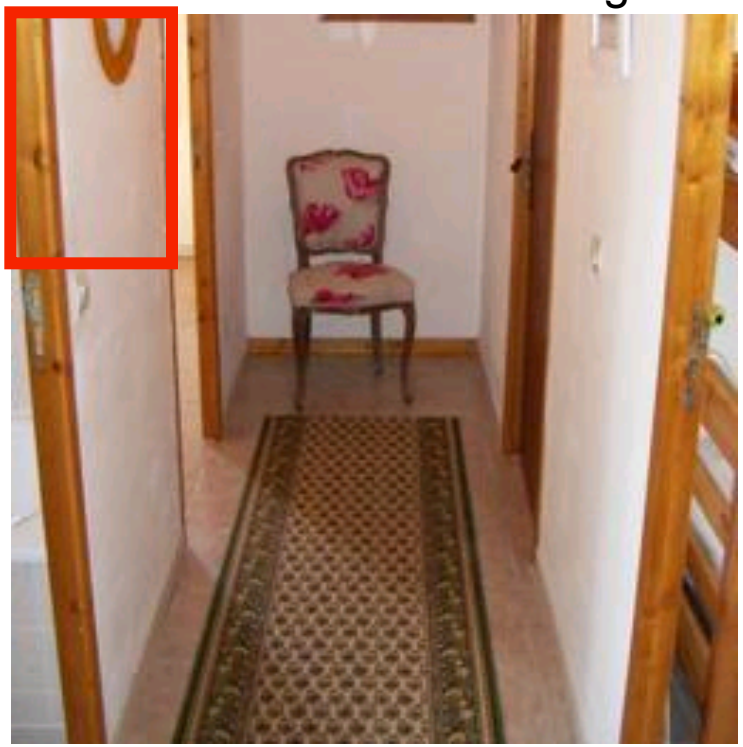


Object recognition
Is it really so hard?

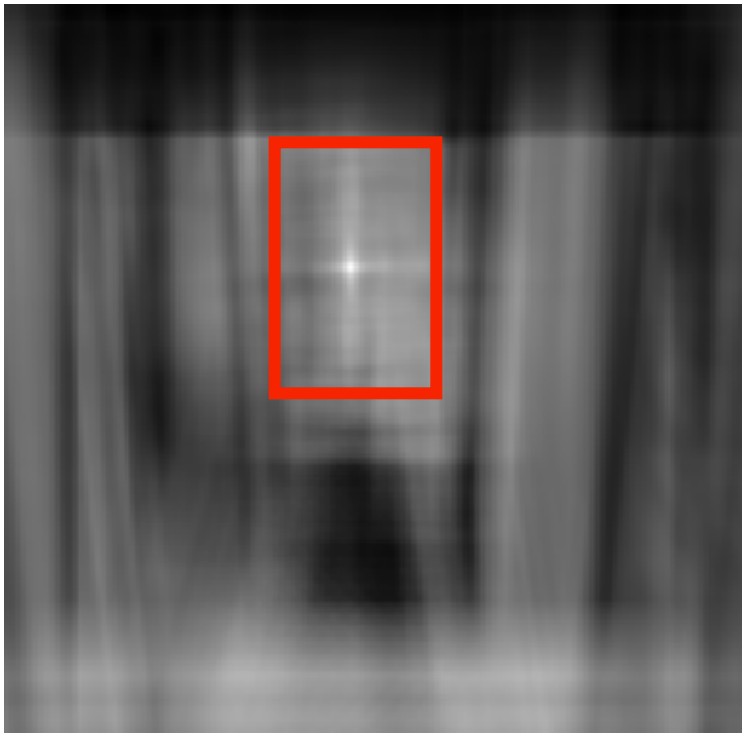
This is a chair

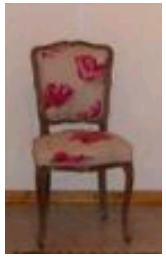


Find the chair in this image



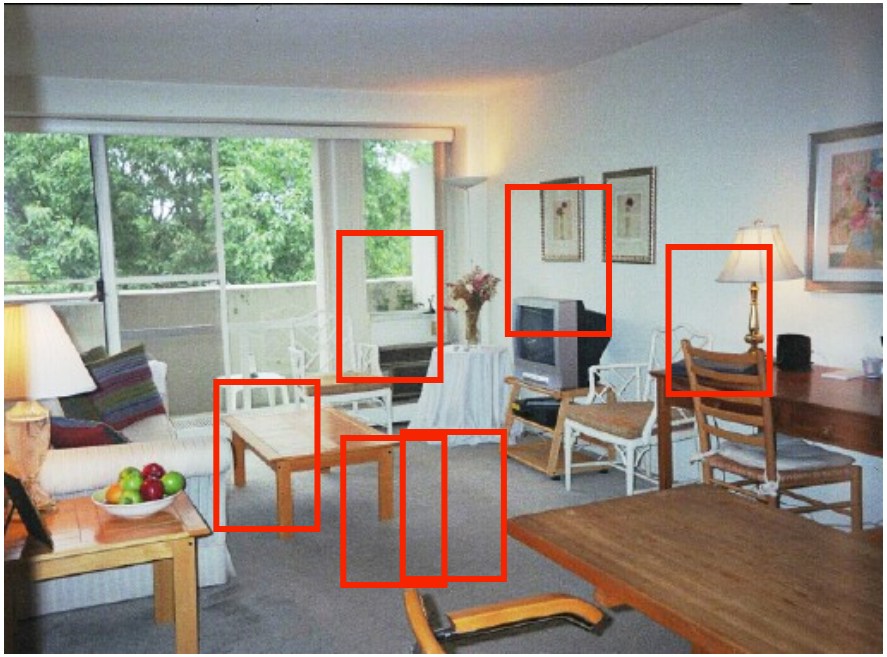
Output of normalized correlation

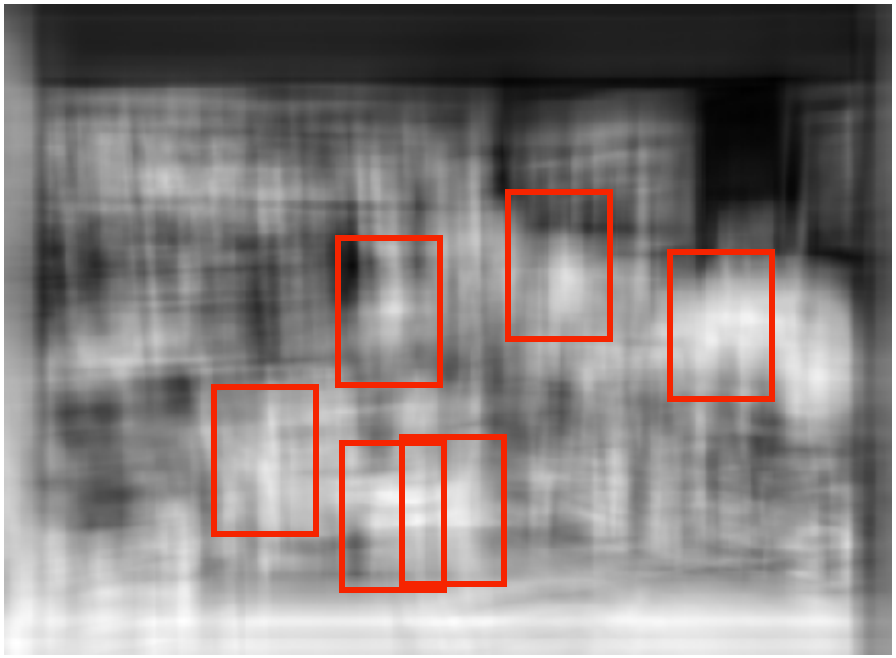




Object recognition
Is it really so hard?

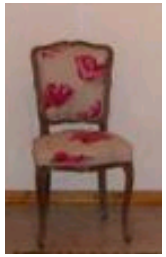
Find the chair in this image





Pretty much garbage
Simple template matching is not going to make it

My biggest concern while making this slide was:
how do I justify 50 years of research, and this course, if this experiment did work?



Object recognition
Is it really so hard?

Find the chair in this image

A “popular method is that of template matching, by point to point correlation of a model pattern with the image pattern. These techniques are inadequate for three-dimensional scene analysis for many reasons, such as occlusion, changes in viewing angle, and articulation of parts.” Nivatia & Binford, 1977.

Why is object recognition a hard task?

Challenges 1: view point variation

Michelangelo 1475-1564

Slides: course object recognition
ICCV 2005

Challenges 2: illumination

slide credit: S. Ullman

Challenges 3: occlusion

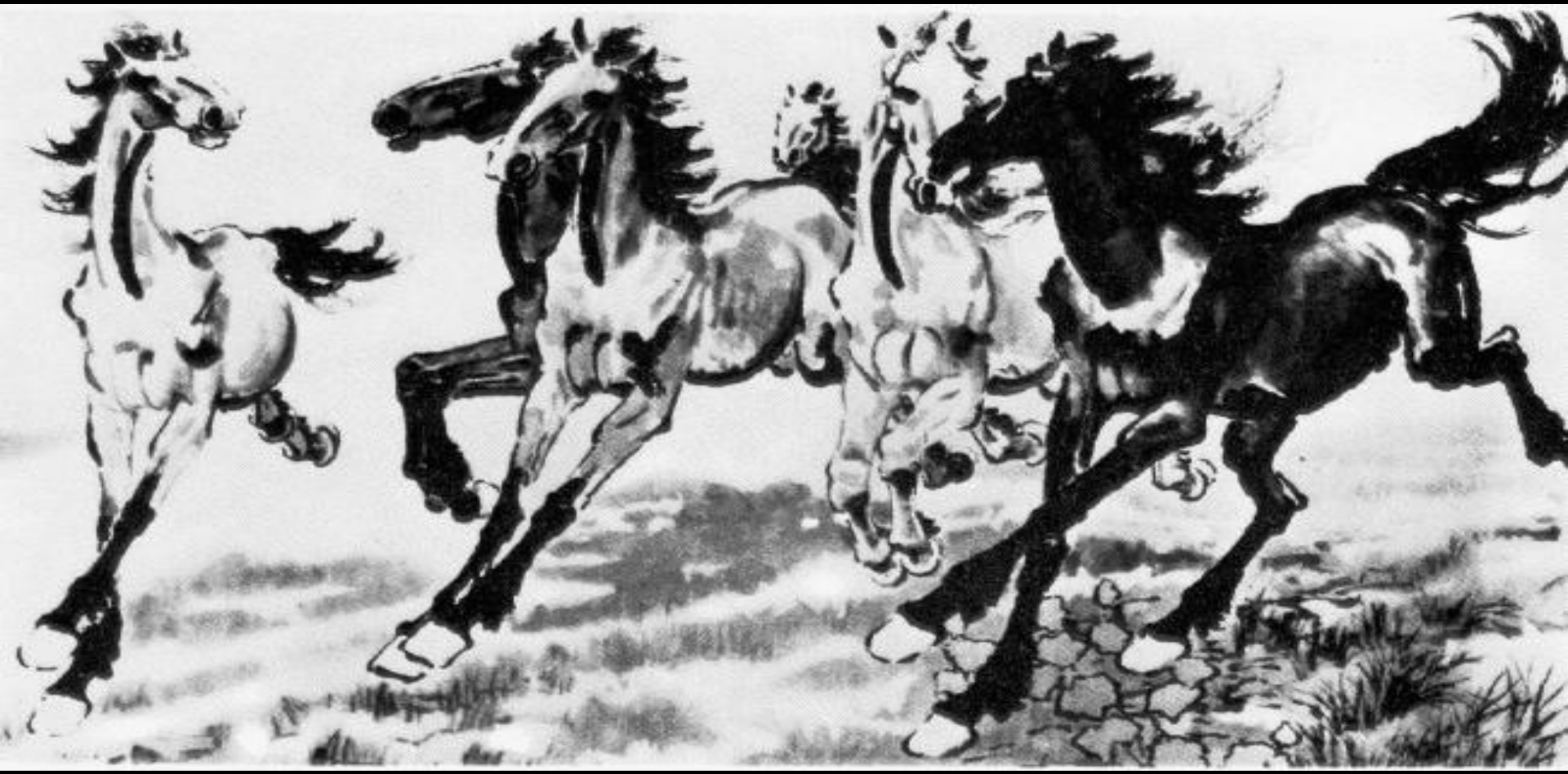
Magritte, 1957

Slides: course object recognition
ICCV 2005

Challenges 4: scale

Slides: course object recognition
ICCV 2005

Challenges 5: deformation



Slides: course object recognition
ICCV 2005

Xu, Beihong 1943

Challenges 6: intra-class variation

Slides: course object recognition
ICCV 2005

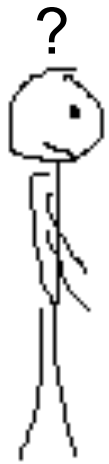
Challenges 7: background clutter

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. J Vis, 3(6), 413-422

Which level of categorization is the right one?

Car is an object composed of:

a few doors, four wheels (not all visible at all times), a roof, front lights, windshield



If you are thinking in buying a car, you might want to be a bit more specific about your categorization.

Entry-level categories
(Jolicoeur, Gluck, Kosslyn 1984)

- Typical member of a basic-level category are categorized at the expected level
- Atypical members tend to be classified at a subordinate level.

A bird

An ostrich

Creation of new categories

A new class can borrow information from similar categories

Object recognition
Is it really so hard?

Yes, object recognition is hard…

(or at least it seems so for now…)

So, let's make the problem simpler: Block world

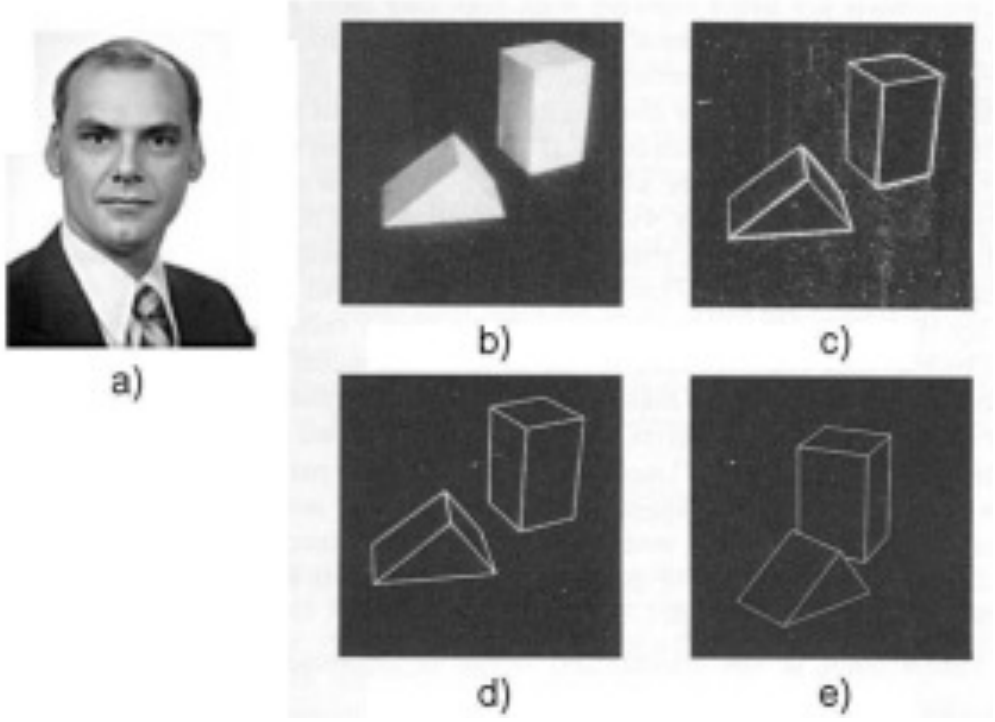


Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - e) are taken from [64] with permission MIT Press.)

Nice framework to develop fancy math, but too far from reality…

Object Recognition in the Geometric Era: a Retrospective. Joseph L. Mundy. 2006

Binford and generalized cylinders

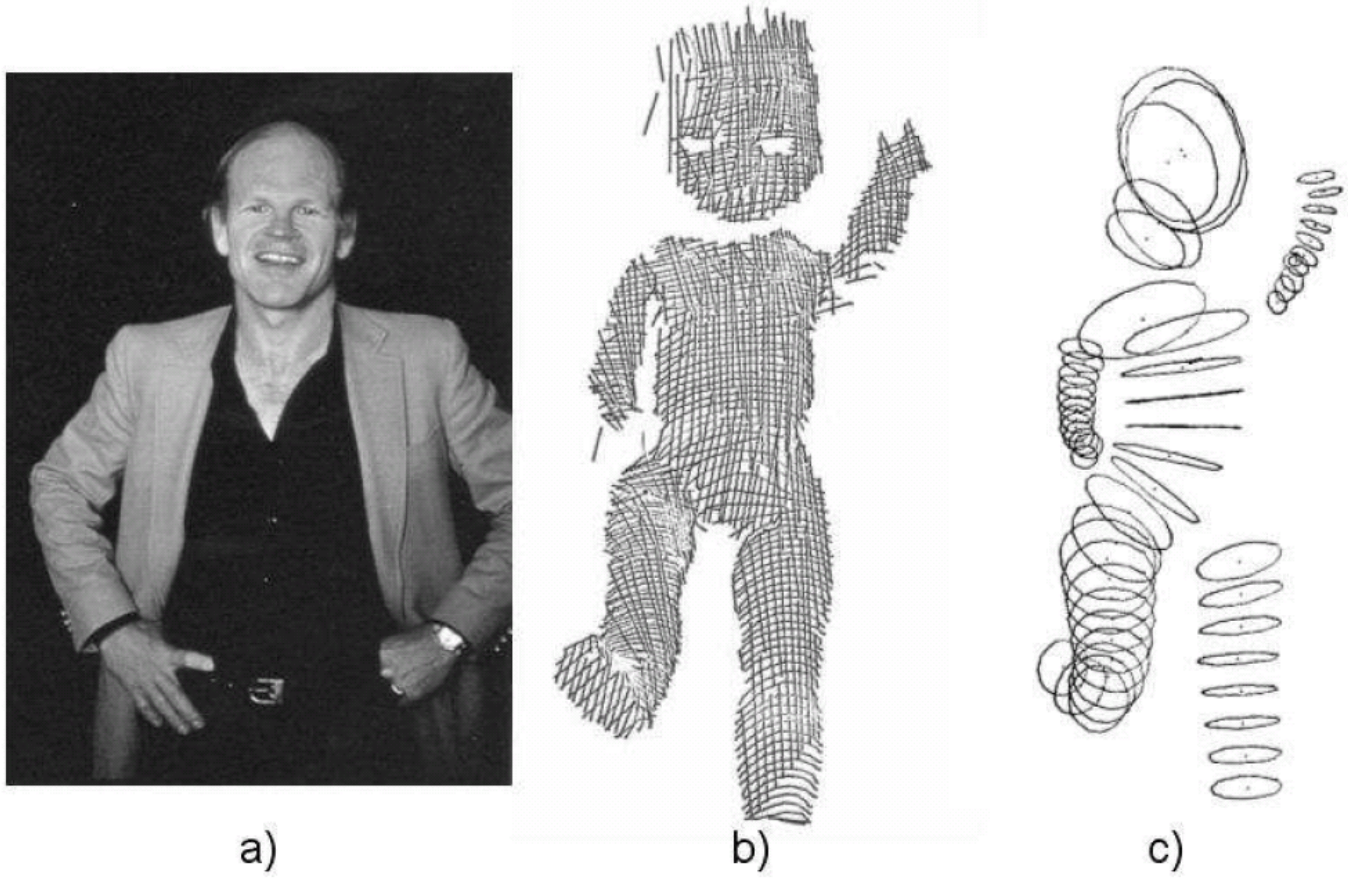
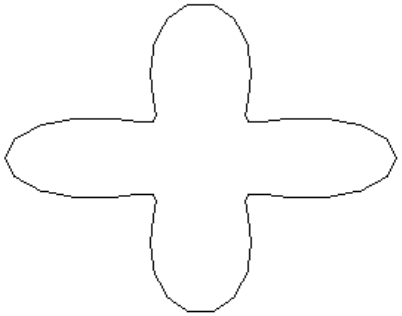


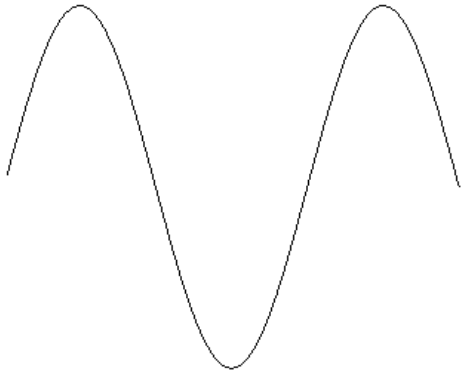
Fig. 3. The representation of objects by assemblies of generalized cylinders. a) Thomas Binford. b) A range image of a doll. c) The resulting set of generalized cylinders. (b) and c) are taken from Agin [1] with permission.)

Object Recognition in the Geometric Era: a Retrospective. Joseph L. Mundy. 2006

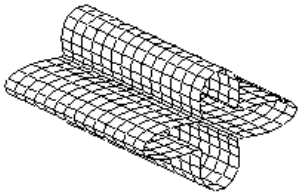
Binford and generalized cylinders



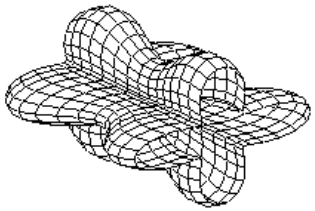
(a) Cross section.



(b) Sweeping rule.



(c) True cylinder



(d) Generalized cylinder

Recognition by components

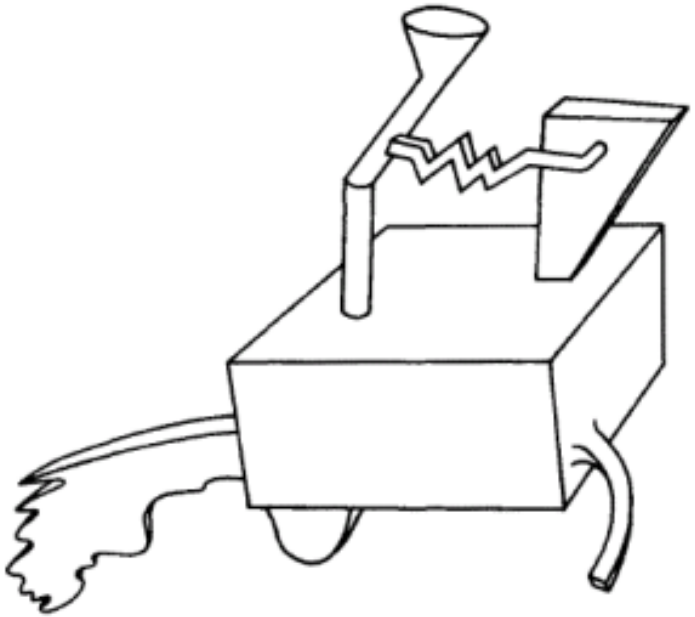
Irving Biederman
Recognition-by-Components: A Theory of Human Image Understanding.
Psychological Review, 1987.

Recognition by components

The fundamental assumption of the proposed theory, recognition-by-components (RBC), is that a modest set of generalized-cone components, called geons *(N* = 36), can be derived from contrasts of five readily detectable properties of edges in a two-dimensional image: curvature, collinearity, symmetry, parallelism, and cotermination.

The “contribution lies in its proposal for a particular vocabulary of components derived from perceptual mechanisms and its account of how an arrangement of these components can access a representation of an object in memory.”

A do-it-yourself example



1) We know that this object is nothing we know
2) We can split this objects into parts that everybody will agree
3) We can see how it resembles something familiar: “a hot dog cart”

“The naive realism that emerges in descriptions of nonsense objects may be reflecting the workings of a representational system by which objects are identified.”

Hypothesis

- Hypothesis: there is a small number of geometric components that constitute the primitive elements of the object recognition system (like letters to form words).

- “The particular properties of edges that are postulated to be relevant to the generation of the volumetric primitives have the desirable properties that they are invariant over changes in orientation and can be determined from just a few points on each edge.”

- Limitation: “The modeling has been limited to concrete entities with specified boundaries.” (count nouns) – this limitation is shared by many modern object detection algorithms.

Constraints on possible models of recognition

1) Access to the mental representation of an object should not be dependent on absolute judgments of quantitative detail

2) The information that is the basis of recognition should be relatively invariant with respect to orientation and modest degradation.

3) Partial matches should be computable. A theory of object interpretation should have some principled means for computing a match for occluded, partial, or new exemplars of a given category.

Stages of processing

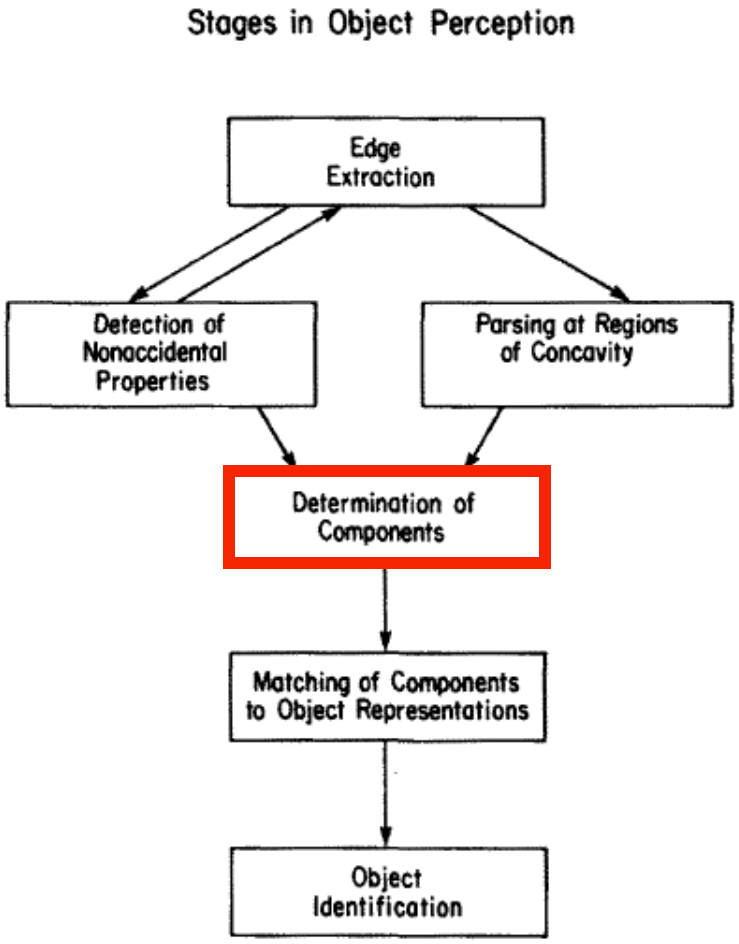


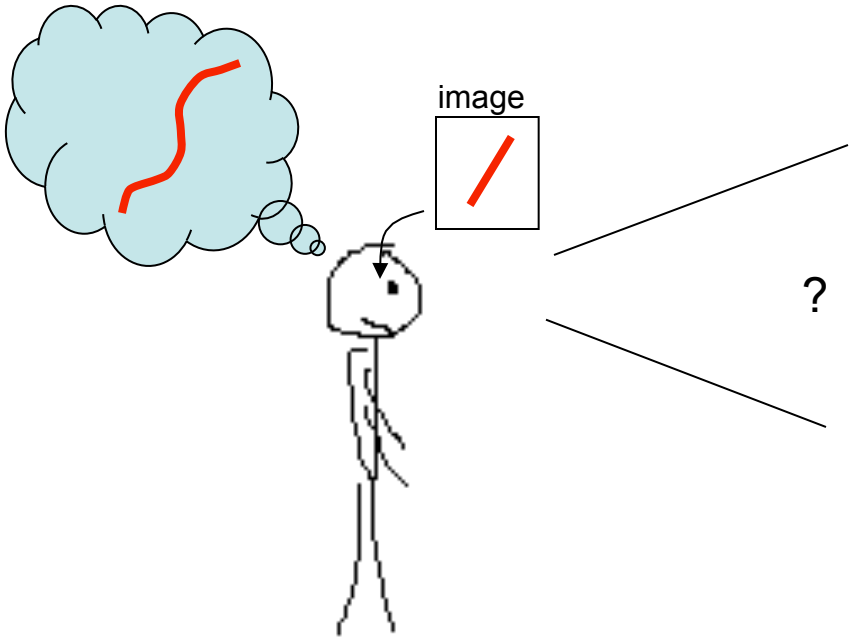
Figure 2. Presumed processing stages in object recognition.

"Parsing is performed, primarily at concave regions, simultaneously with a detection of nonaccidental properties."

Non accidental properties

Certain properties of edges in a two-dimensional image are taken by the visual system as strong evidence that the edges in the three-dimensional world contain those same properties.

Non accidental properties, (Witkin & Tenenbaum,1983): Rarely be produced by accidental alignments of viewpoint and object features and consequently are generally unaffected by slight variations in viewpoint.



Examples:

- Colinearity
- Smoothness
- Symmetry
- Parallelism
- Cotermination

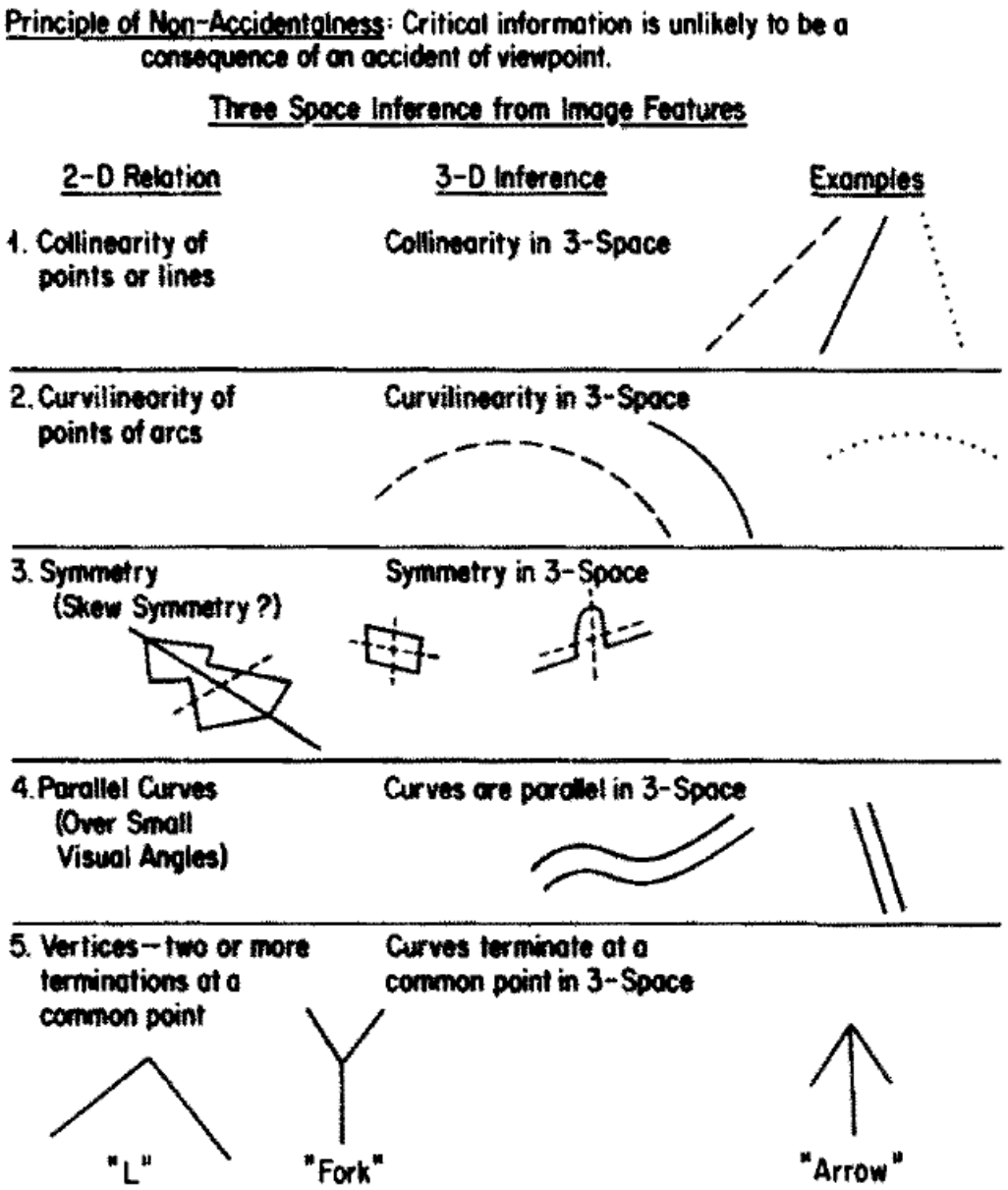
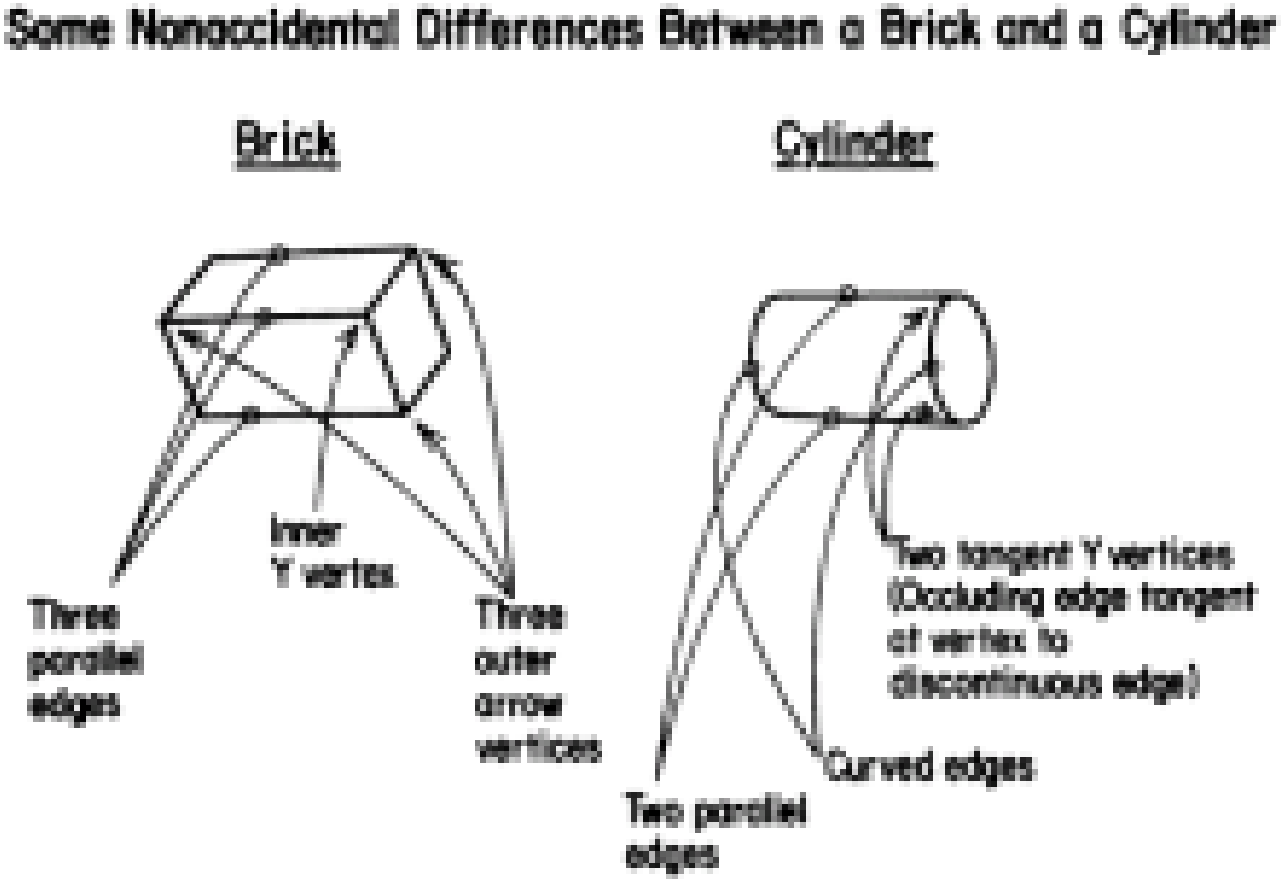
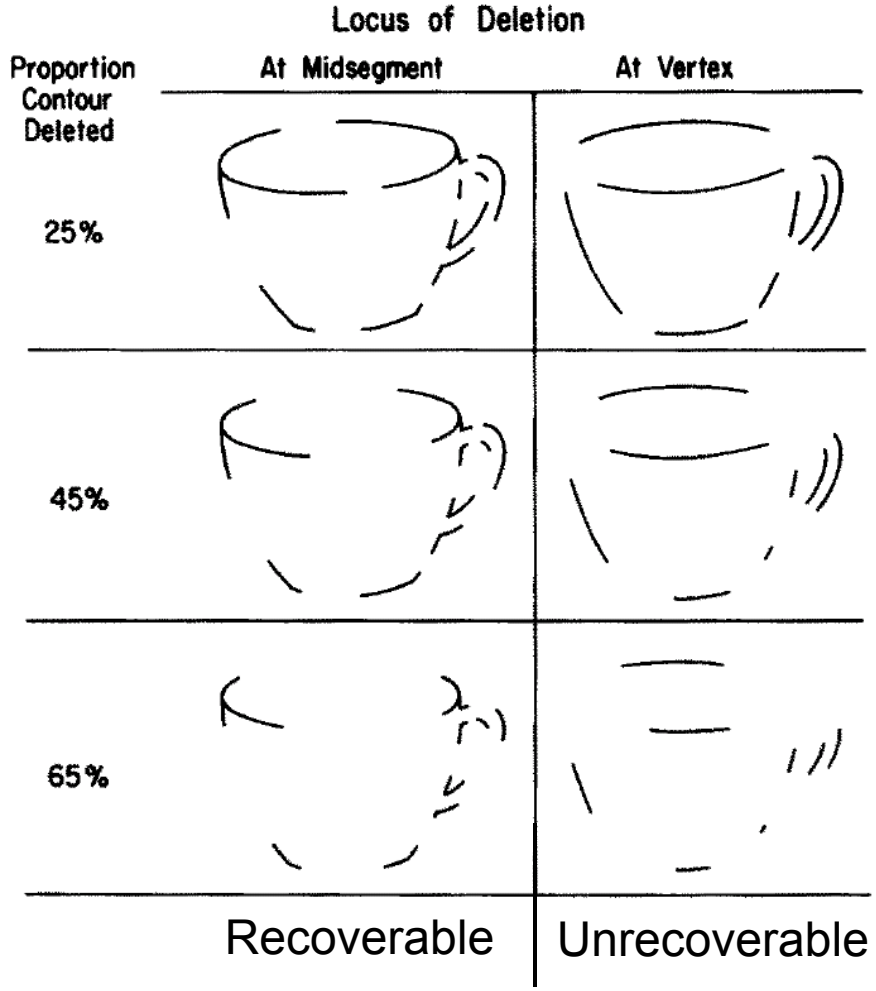


Figure 4. Five nonaccidental relations. (From Figure 5.2, *Perceptual organization and visual recognition* [p. 77] by David Lowe. Unpublished doctorial dissertation, Stanford University. Adapted by permission.)



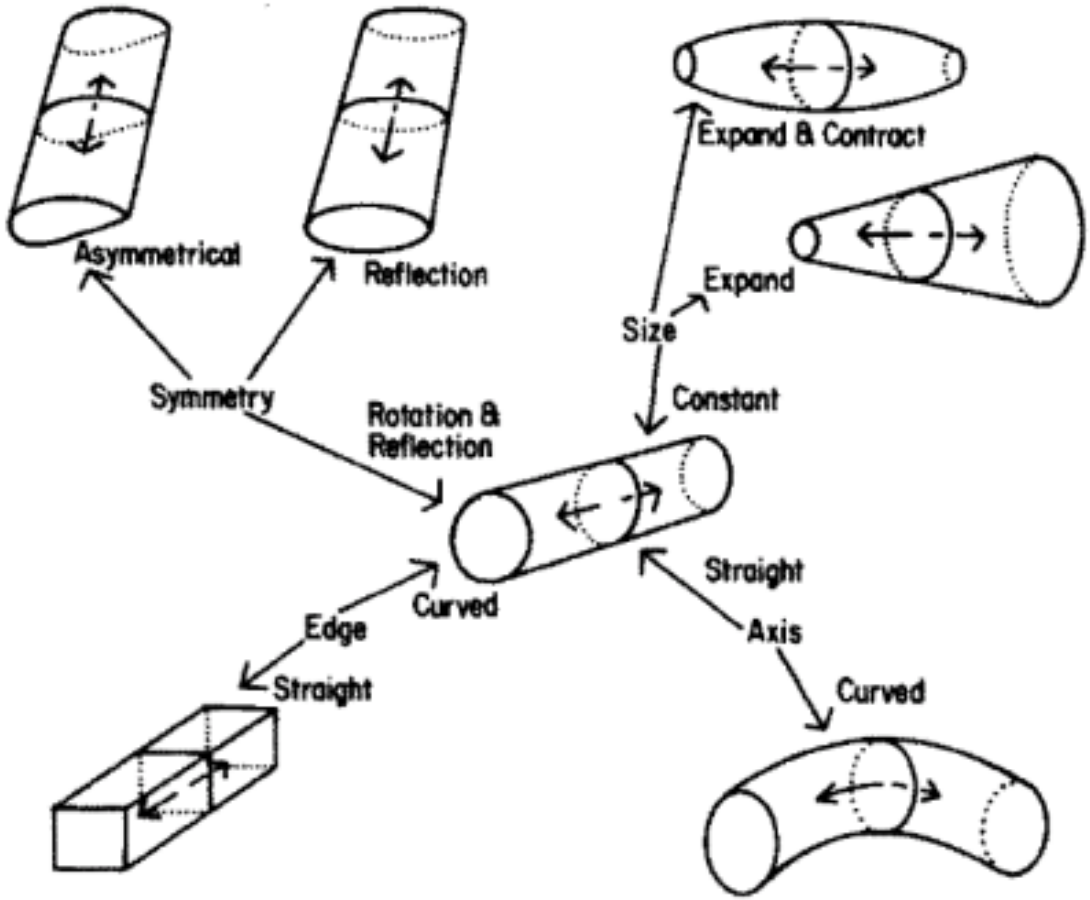
The high speed and accuracy of determining a given nonaccidental relation {e.g., whether some pattern is symmetrical) should be contrasted with performance in making absolute quantitative judgments of variations in a single physical attribute, such as length of a segment or degree of tilt or curvature.

Object recognition is performed by humans in around 100ms.



“If contours are deleted at a vertex they can be restored, as long as there is no accidental filling-in. The greater disruption from vertex deletion is expected on the basis of their importance as diagnostic image features for the components.”

From generalized cylinders to GEONS



"From variation over only two or three levels in the nonaccidental relations of four attributes of generalized cylinders, a set of 36 GEONS can be generated."

Geons represent a restricted form of generalized cylinders.

More GEONS

CROSS SECTION

Geon	Edge Straight S Curved C	Symmetry Rot & Ref ++ Ref + Asymm –	Size Constant ++ Expanded – Exp & Cont – –	Axis Straight + Curved –
	S	+ +	+ +	+
	C	+ +	+ +	+
	S	+	–	+
	S	+ +	+	–
	C	+ +	–	+
	S	+	+	+

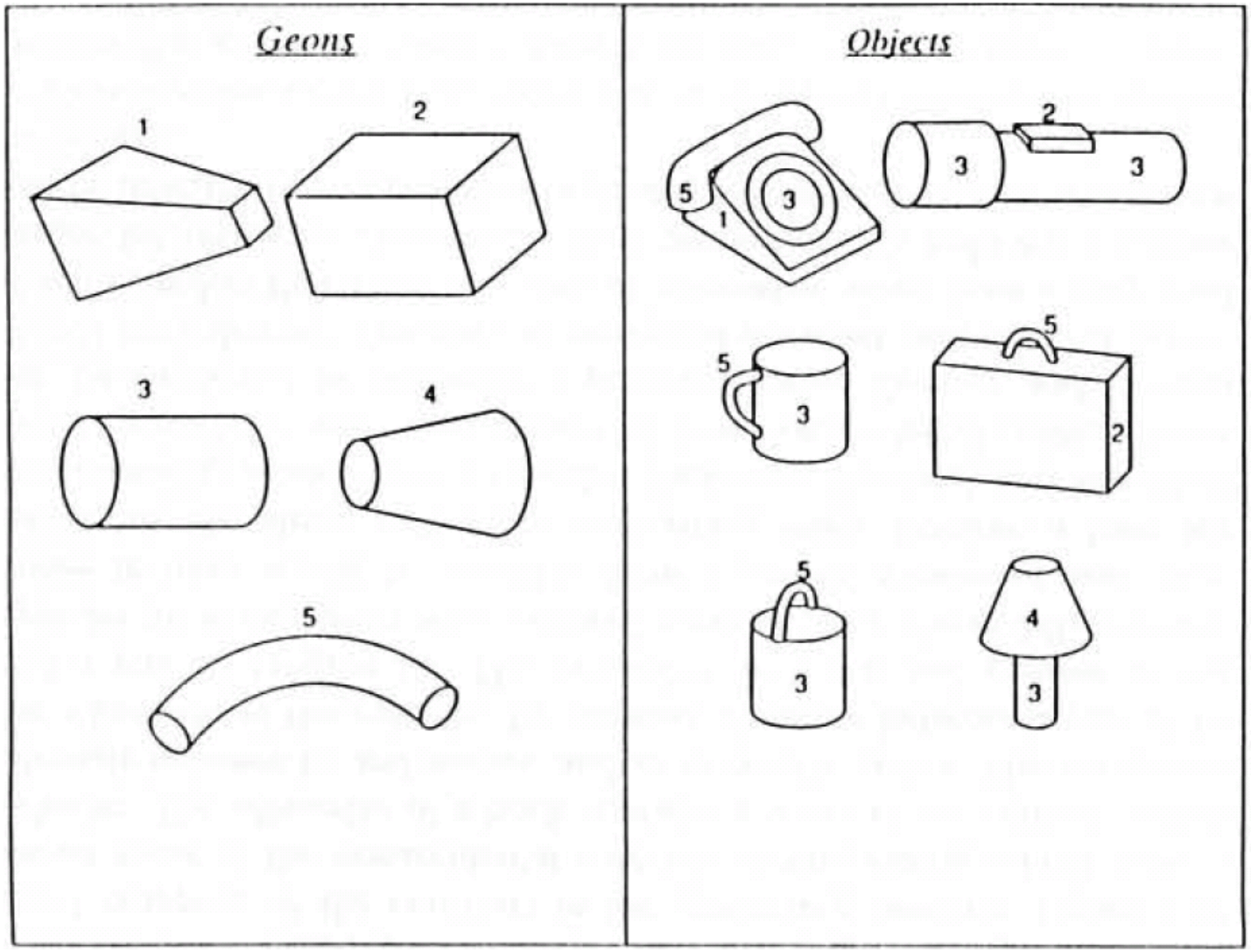
Figure 7. Proposed partial set of volumetric primitives (geons) derived from differences in nonaccidental properties.

CROSS SECTION

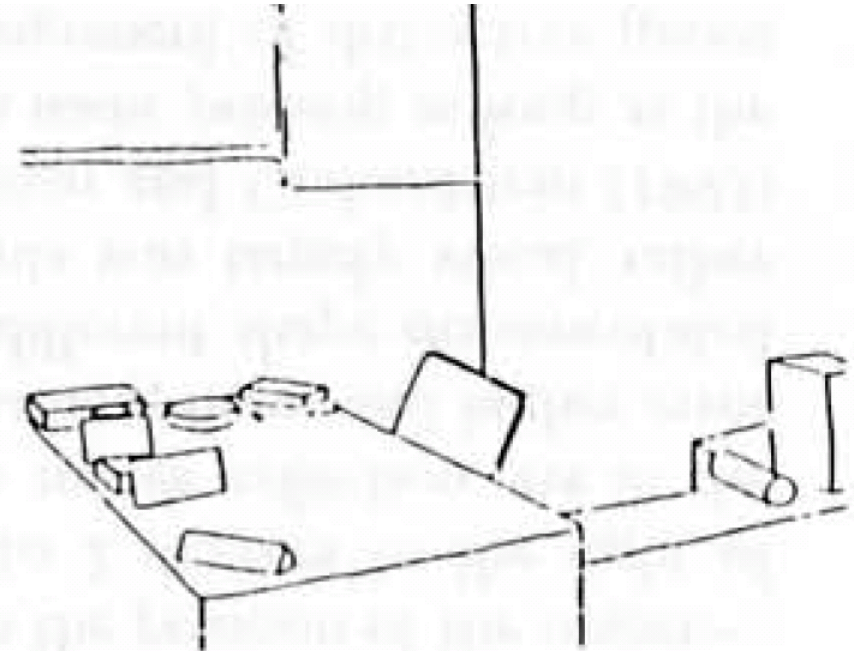
Geon	Edge Straight S Curved C	Symmetry Rot & Ref ++ Ref + Asymm –	Size Constant ++ Expanded – Exp & Cont – –	Axis Straight + Curved –
	S	+	+ +	–
	C	+	+ +	–
	S	+ +	–	–
	C	+ +	–	–
	S	+	–	–
	C	+	–	–

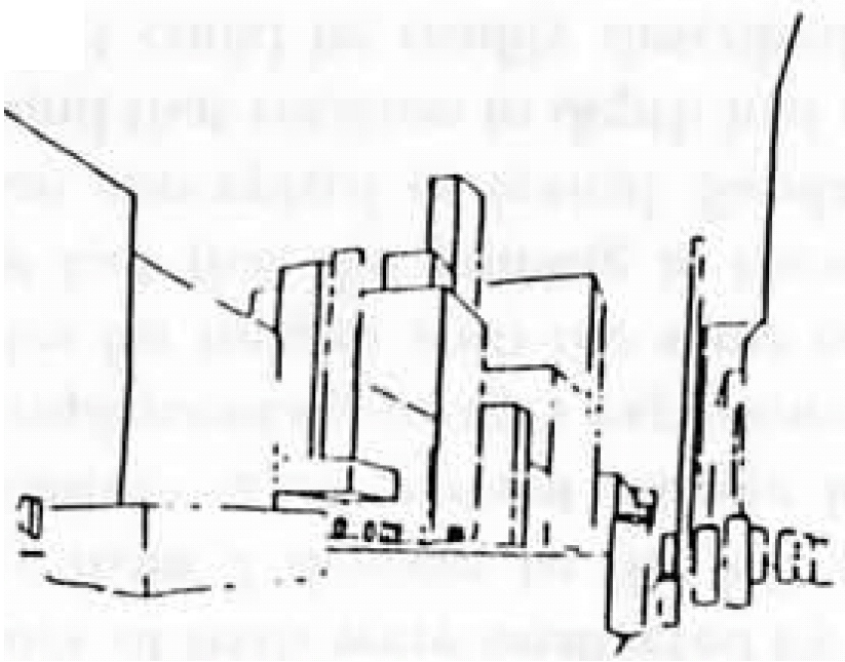
Figure 9. Geons with curved axis and straight or curved cross sections. (Determining the shape of the cross section, particularly if straight, might require attention.)

Objects and their geons



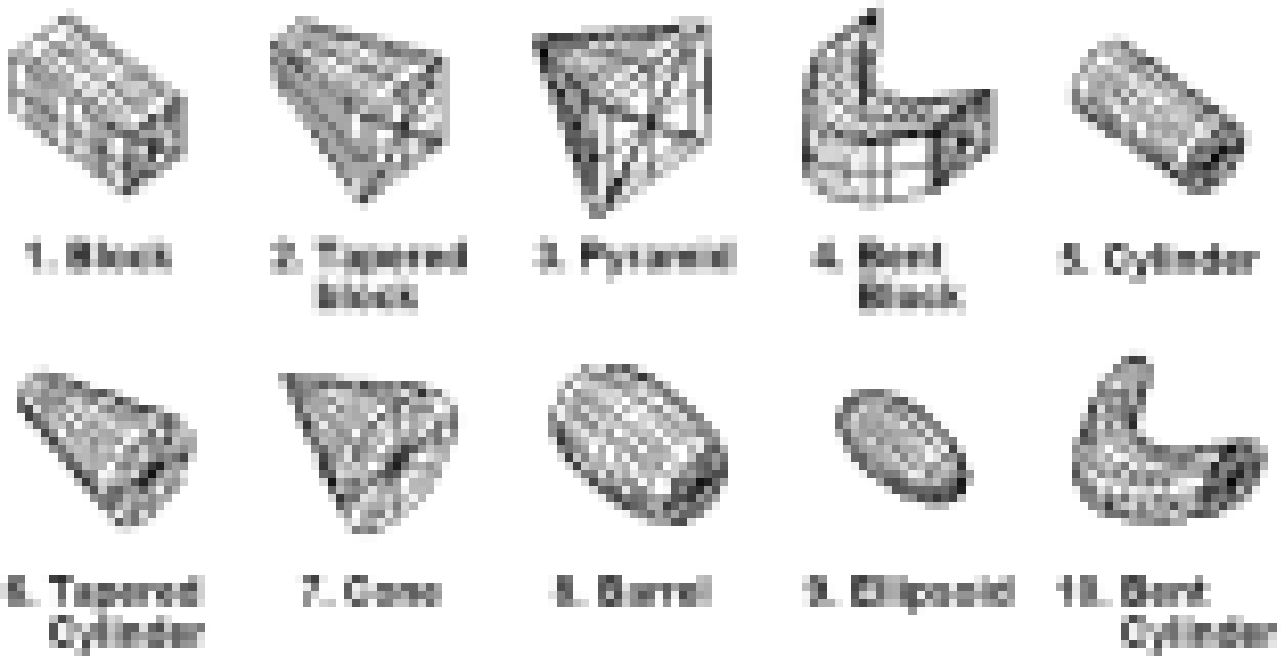
Scenes and geons





Mezzanotte & Biederman

Supercuadrics



Introduced in computer vision by A. Pentland, 1986.

What is missing?

The notion of geometric structure.

Although they were aware of it, the previous works put more emphasis on defining the primitive elements than modeling their geometric relationships.

The importance of spatial arrangement

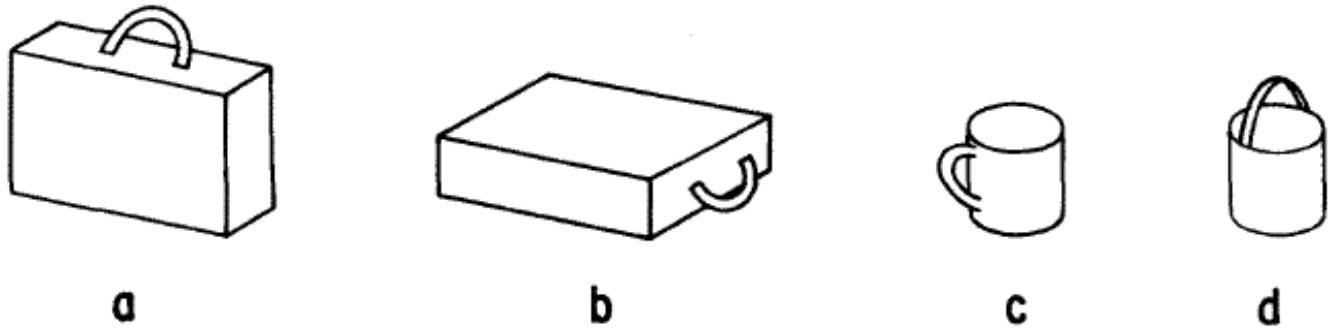


Figure 3. Different arrangements of the same components can produce different objects.

Parts and Structure approaches

With a different perspective, these models focused more on the geometry than on defining the constituent elements:

- Fischler & Elschlager 1973
- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Perona et al. ‘95, ‘96, ’98, ’00, ’03, ‘04
- Felzenszwalb & Huttenlocher ’00, ’04
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ’03, ’04
- Many papers since 2000

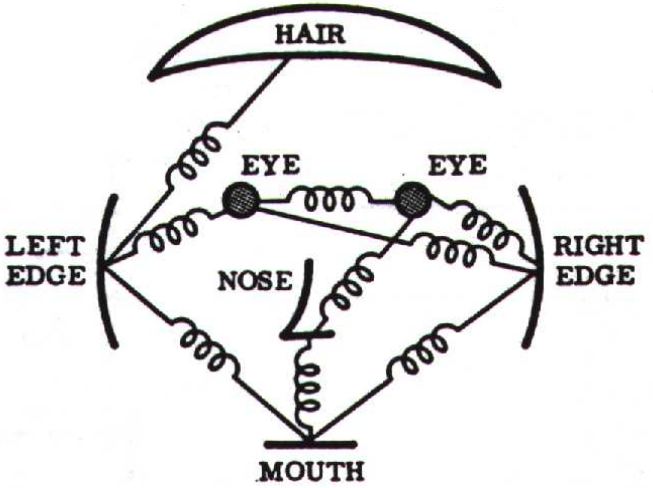


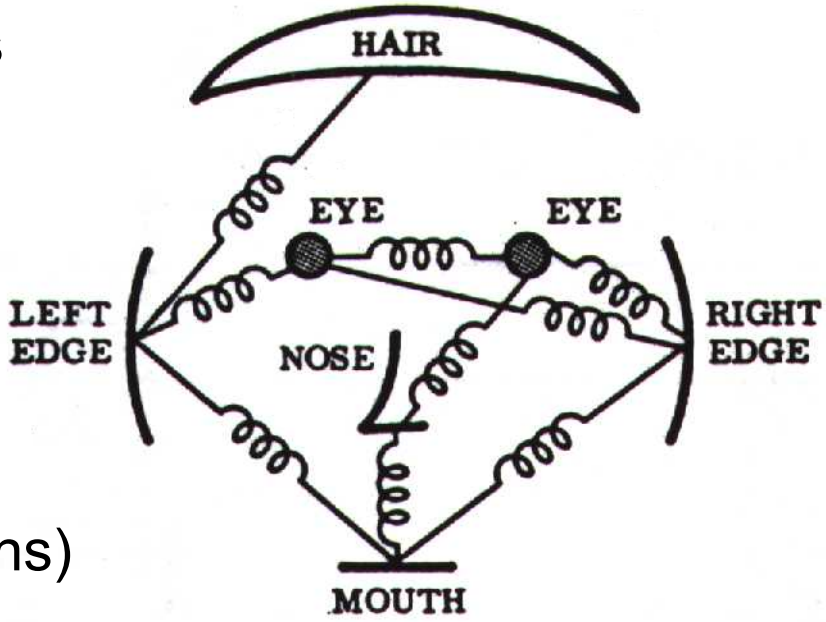
Figure from [Fischler & Elschlager 73]

Representation

- Object as set of parts
  - Generative representation

- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter



We will discuss these models more in depth later

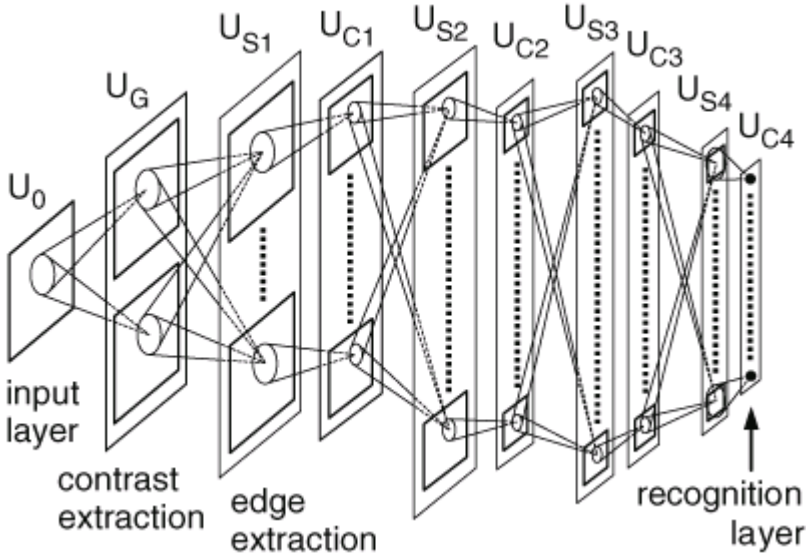
But, despite promising initial results…things did not work out so well (lack of data, processing power, lack of reliable methods for low-level and mid-level vision)

Instead, a different way of thinking about object detection started making some progress: learning based approaches and classifiers, which ignored low and mid-level vision.

Maybe the time is here to come back to some of the earlier models, more grounded in intuitions about visual perception.

Neocognitron

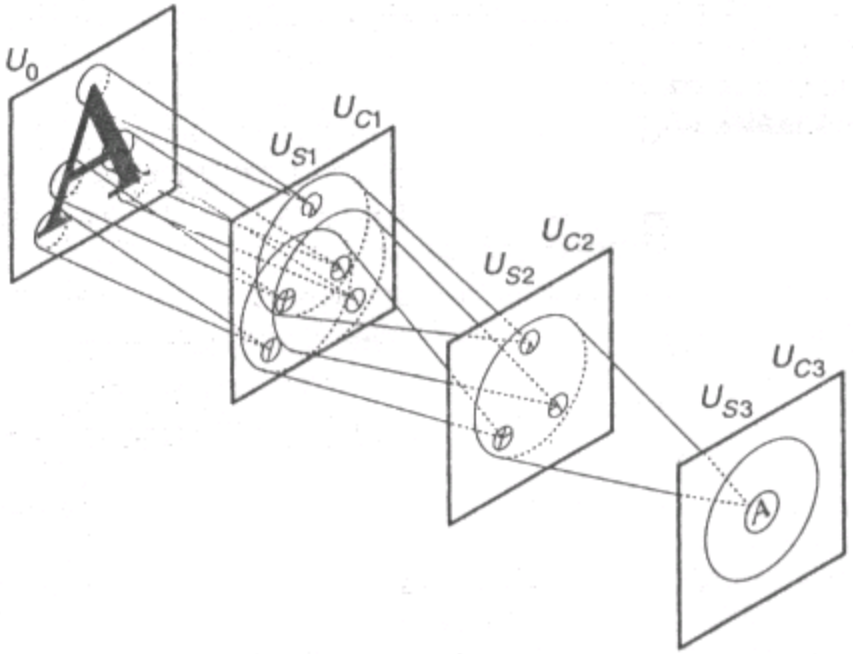
Fukushima (1980). Hierarchical multilayered neural network

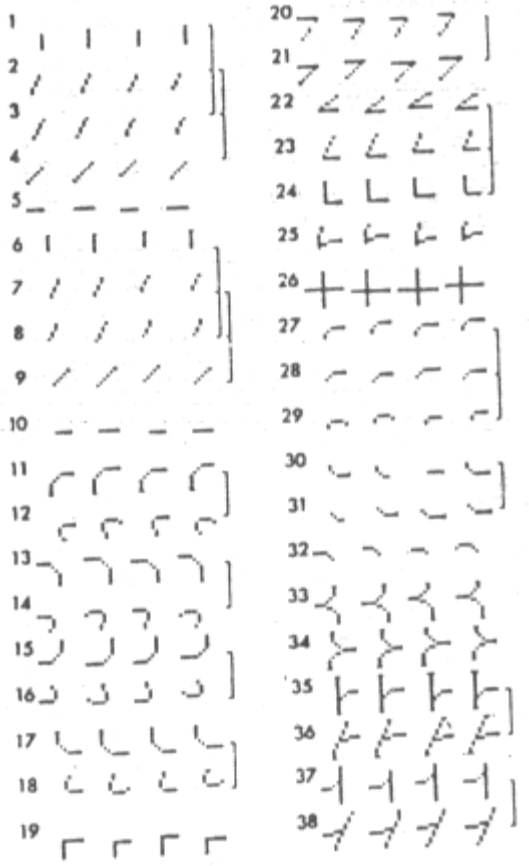


S-cells work as feature-extracting cells. They resemble simple cells of the primary visual cortex in their response.

C-cells, which resembles complex cells in the visual cortex, are inserted in the network to allow for positional errors in the features of the stimulus. The input connections of C-cells, which come from S-cells of the preceding layer, are fixed and invariable. Each C-cell receives excitatory input connections from a group of S-cells that extract the same feature, but from slightly different positions. The C-cell responds if at least one of these S-cells yield an output.

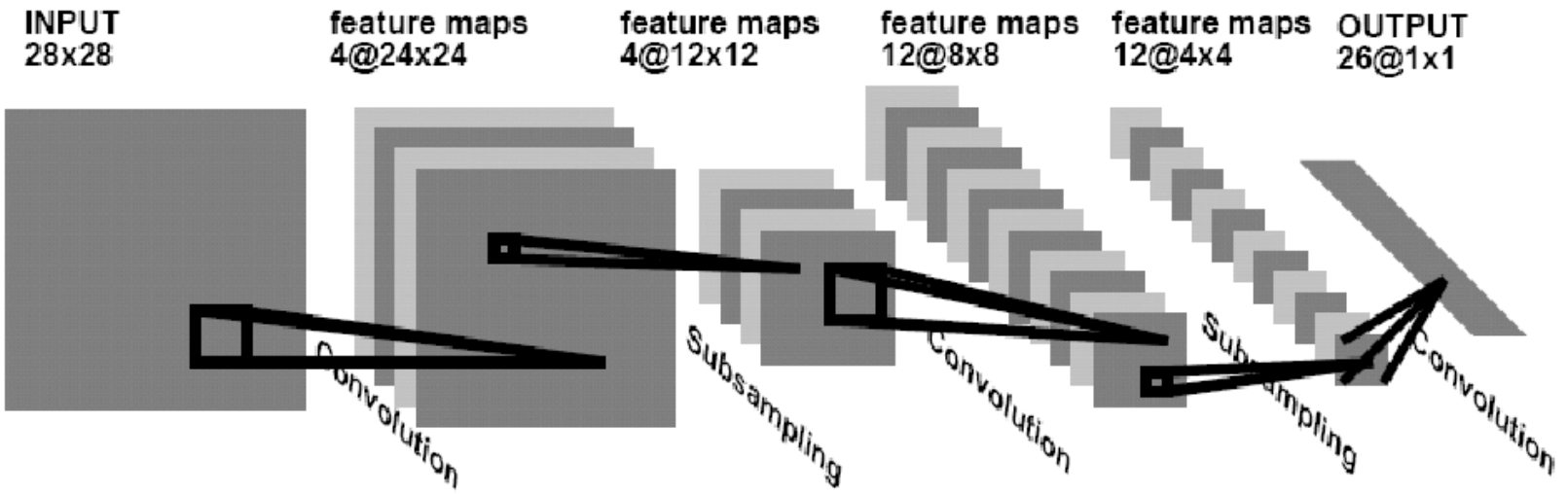
Neocognitron





Learning is done greedily for each layer

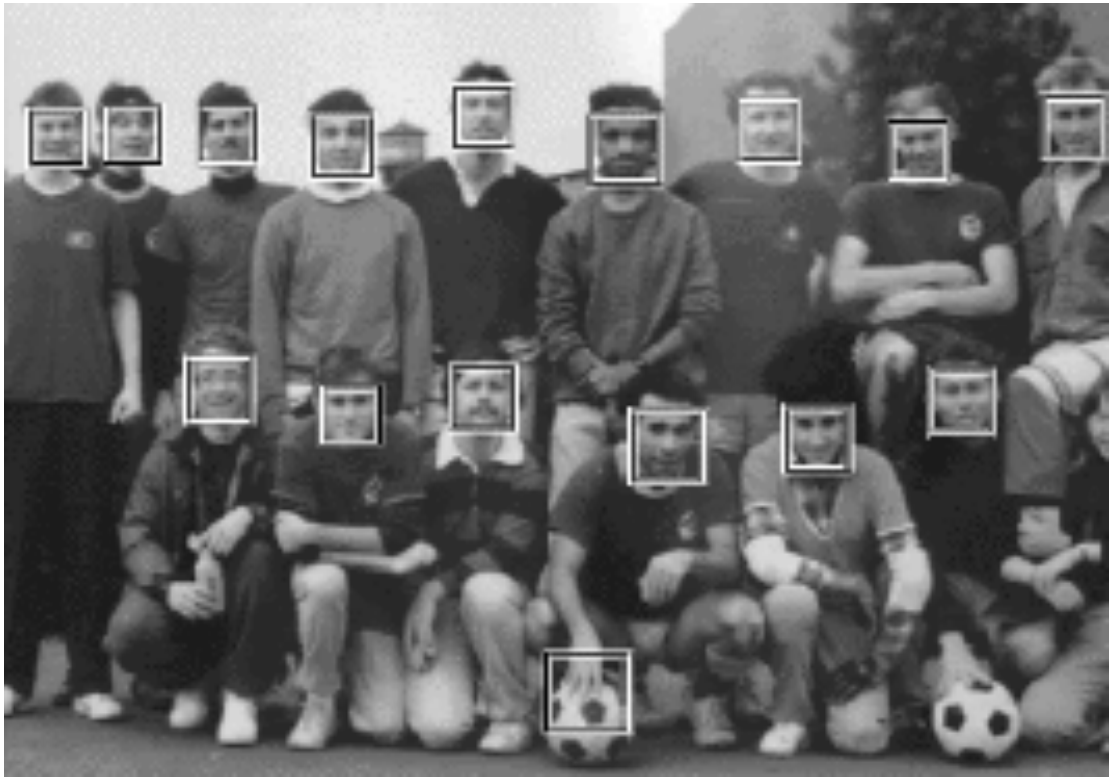
Convolutional Neural Network



Le Cun et al, 98

The output neurons share all the intermediate levels

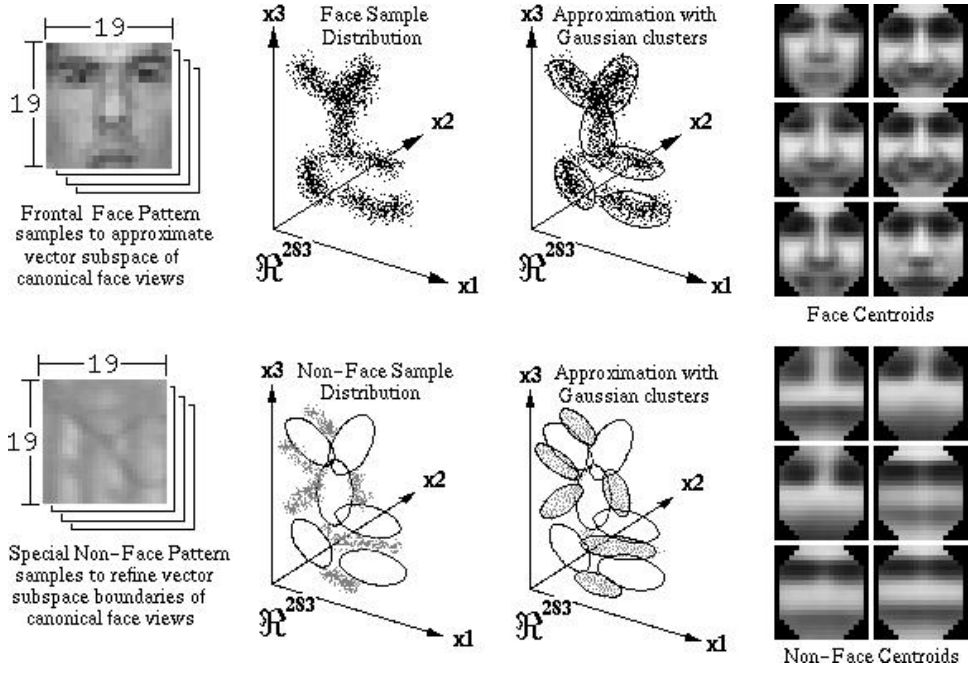
Face detection and the success of learning based approaches

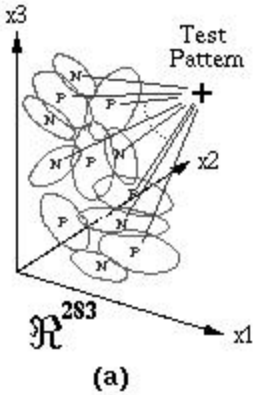


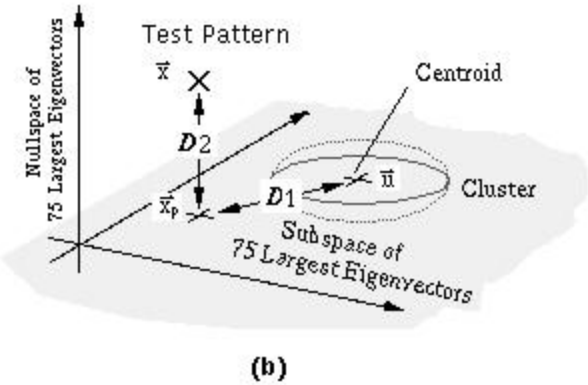
- The representation and matching of pictorial structures Fischler, Elschlager (1973).
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….

Distribution-Based Face Detector

- Learn face and nonface models from examples [Sung and Poggio 95]
- Cluster and project the examples to a lower dimensional space using Gaussian distributions and PCA
- Detect faces using distance metric to face and nonface clusters

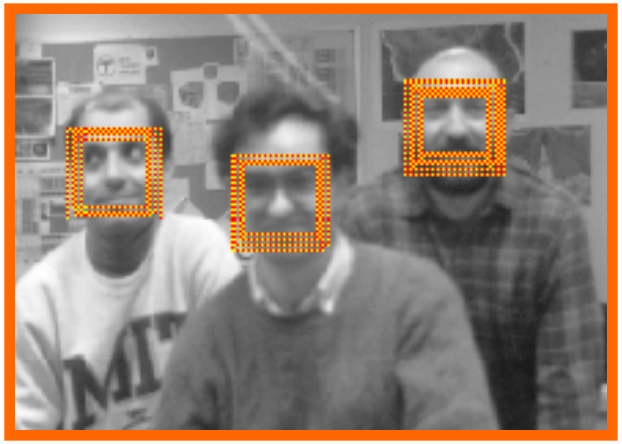






Distribution-Based Face Detector

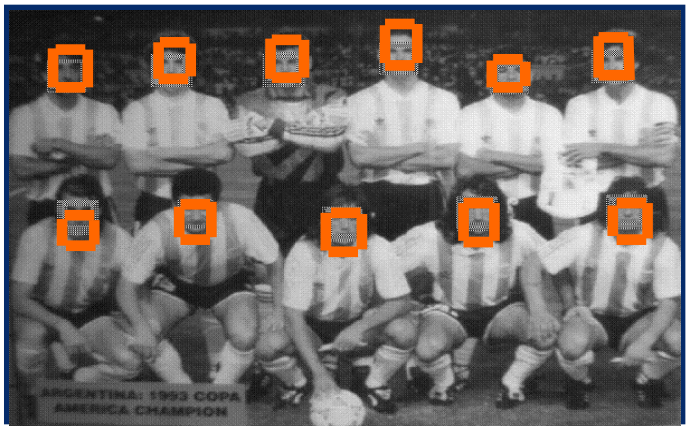
- Learn face and nonface models from examples [Sung and Poggio 95]



Training Database

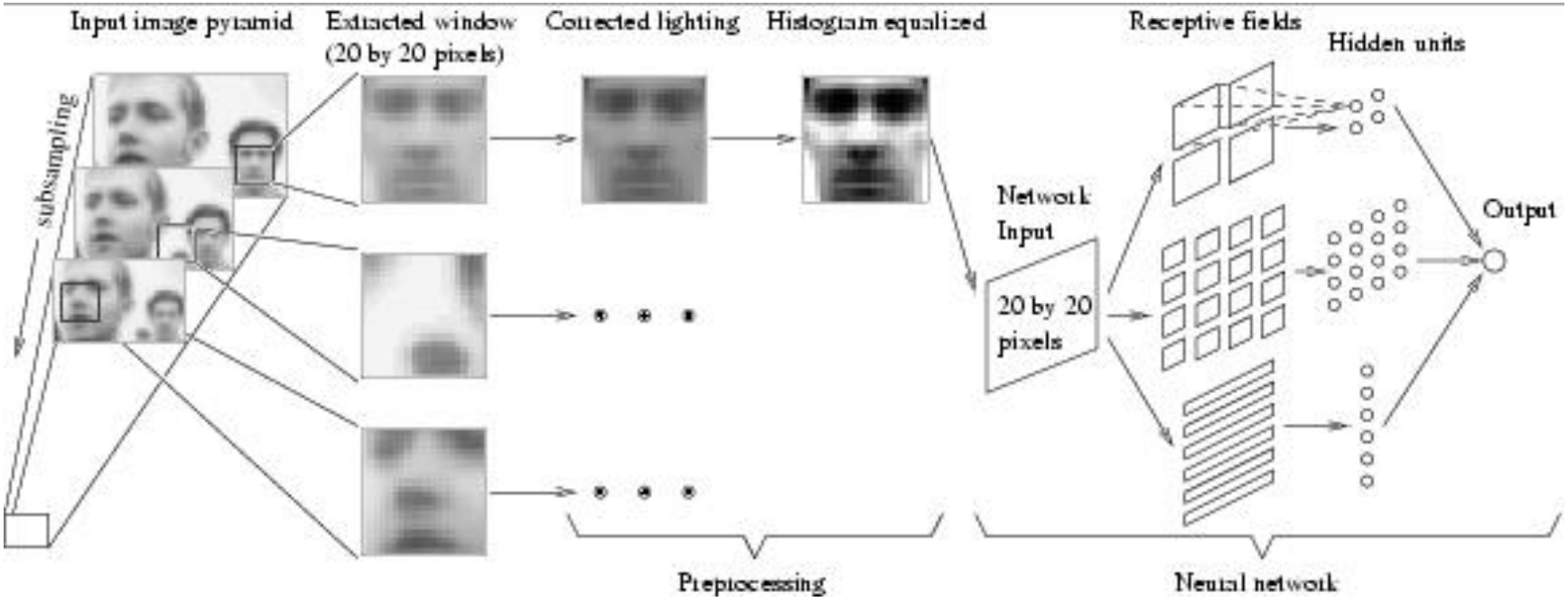
1000+ Real, 3000+ *VIRTUAL*

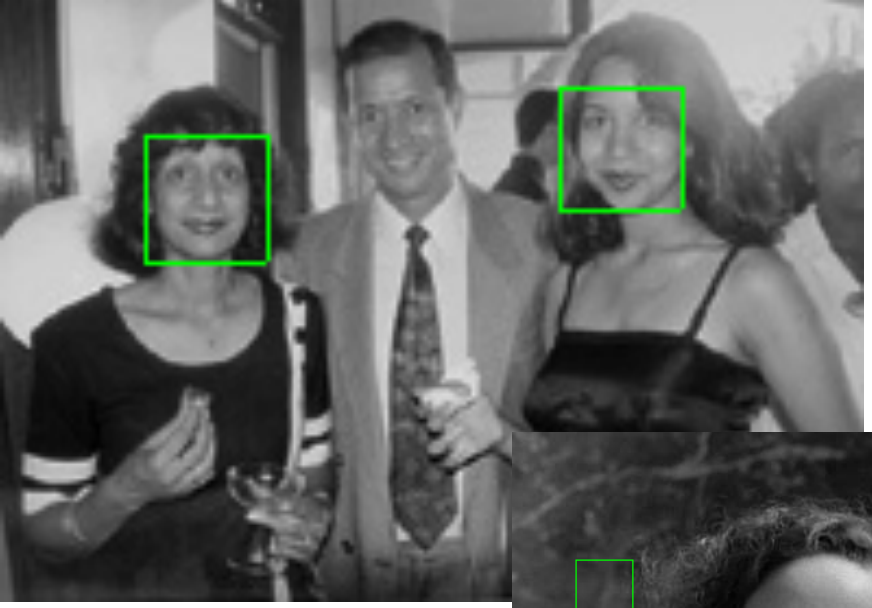
50,0000+ Non-Face Pattern



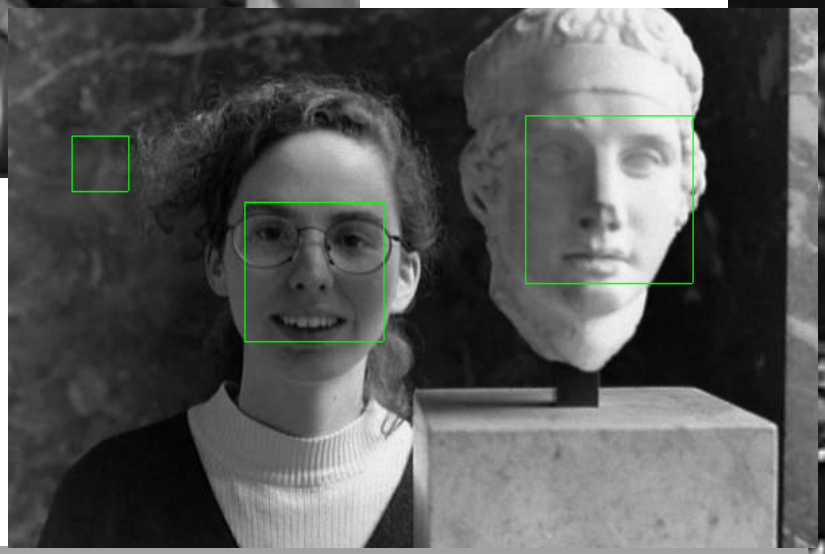
Neural Network-Based Face Detector

- Train a set of multilayer perceptrons and arbitrate a decision among all outputs [Rowley et al. 98]






STAR TREK
DEEP SPACE NINE
THE WAY OF THE WARRIOR
A Novel by Diane Carey



Coarse-to-Fine Face Detection

François Fleuret * Donald Geman †

June 2000

for other objects in various subsets.

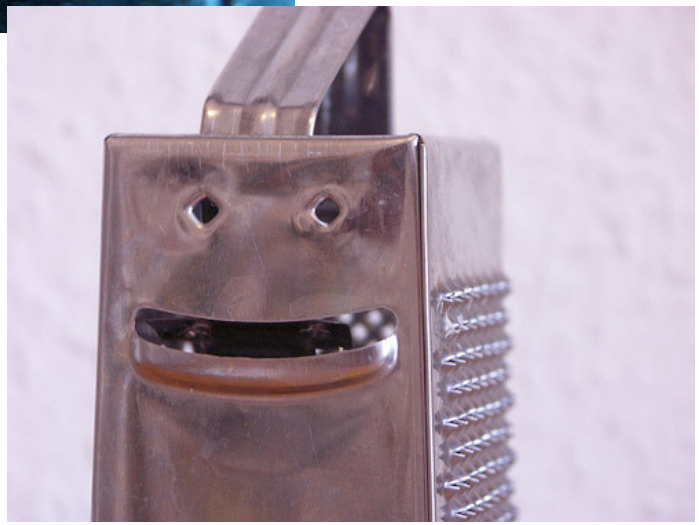
Finally, in defense of limited goals, nobody has yet demonstrated that objects from even one generic class under constrained poses can be rapidly detected without errors in complex, natural scenes; visual selection by humans occurs within two hundred milleseconds and is virtually perfect.

Acknowledgements: We are grateful to Yali Amit for many suggestions during a

*Avant-Projet IMEDIA, INRIA-Rocquencourt, Domaine de Voluceau, B.P.105, 78153 Le Chesnay. Email:Francois.Fleuret@inria.fr. Supported in part by the CNET.

†Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003. Email:geman@math.umass.edu. Supported in part by ONR under contract N00014-97-1-0249 and ARO under MURI grant DAAH04-96-1-0445.

Faces everywhere



http://www.marcofolio.net/imagedump/faces_everywhere_15_images_8_illusions.html

Rapid Object Detection Using a Boosted Cascade of Simple Features

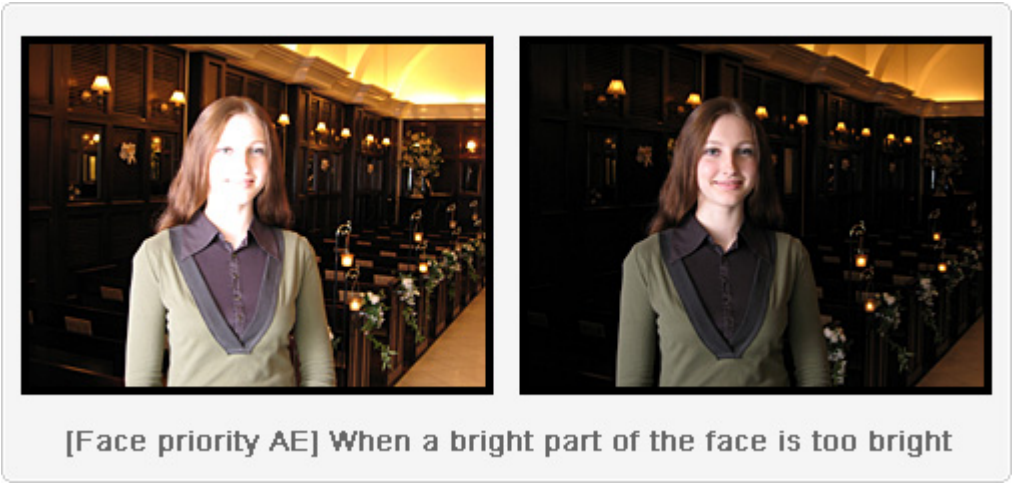
Paul Viola Michael J. Jones
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA

Most of this work was done at Compaq CRL before the authors moved to MERL

Manuscript available on web:

http://citeseer.ist.psu.edu/cache/papers/cs/23183/http:zSzzSzwww.ai.mit.eduzSzpeoplezSzviolazSzresearchzSzpublicationszSzICCV01-Viola-Jones.pdf/viola01robust.pdf

Face detection



[Face priority AE] When a bright part of the face is too bright

Families of recognition algorithms

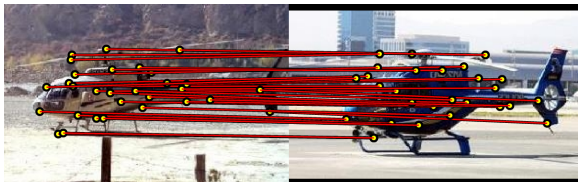
Bag of words models

Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

Voting models

Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schneiderman, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching
Deformable models



Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

Constellation models

Fischler and Elschlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

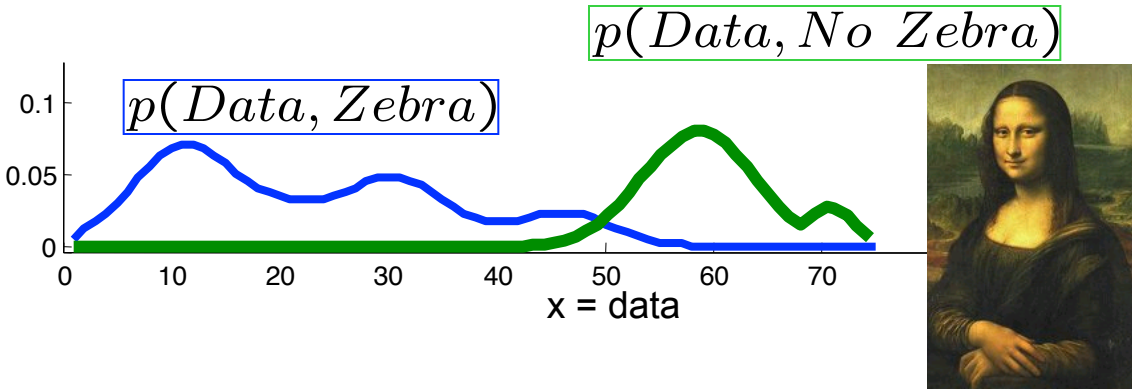
Rigid template models

Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006

Discriminative vs. generative

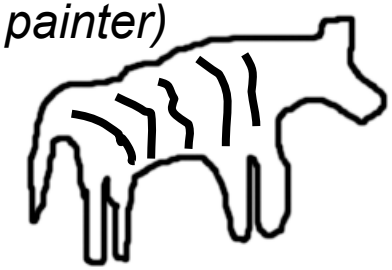
- Generative model

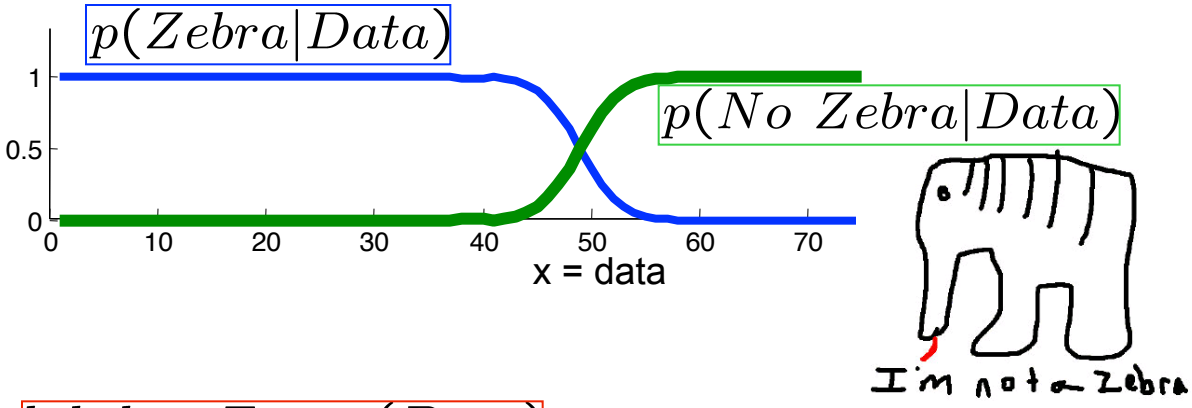
  (*The artist*)



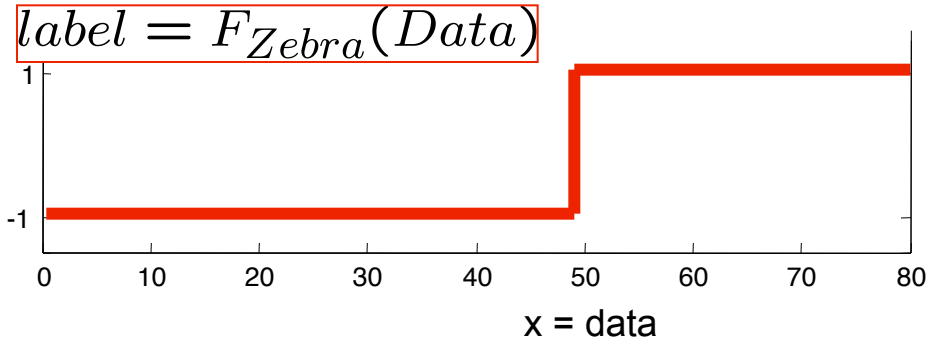
- Discriminative model

(The lousy painter)





- Classification function



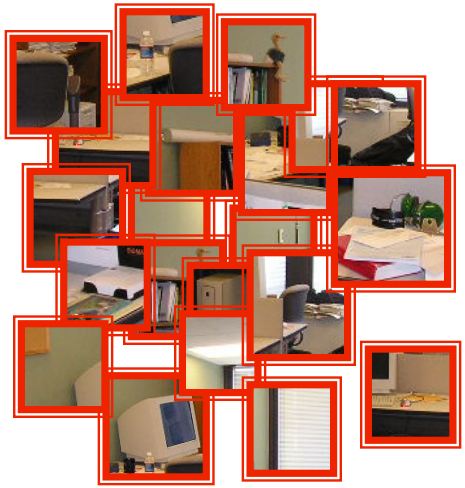
Discriminative methods

Object detection and recognition is formulated as a classification problem.

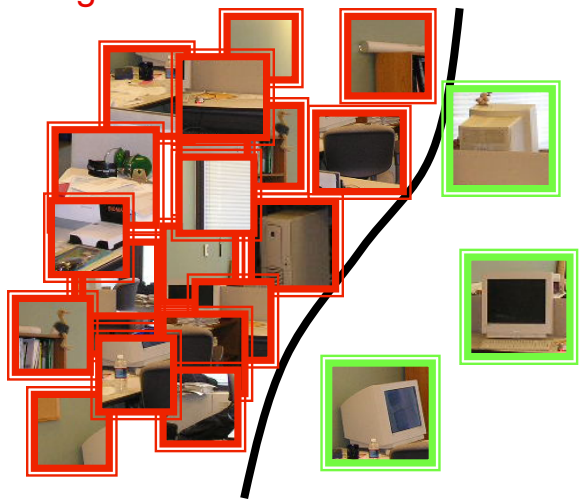
The image is partitioned into a set of overlapping windows

… and a decision is taken at each window about if it contains a target object or not.

Where are the screens?



Bag of image patches



In some feature space

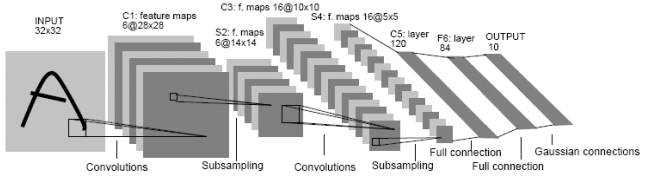
Discriminative methods

Nearest neighbor

10^6 examples

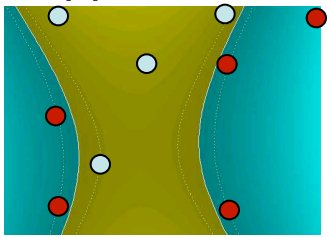
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005
…

Neural networks



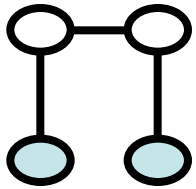
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001
…

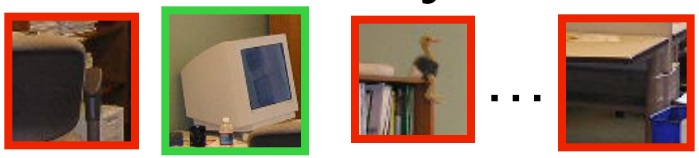
Conditional Random Fields



McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003
…

Formulation

- Formulation: binary classification



Features	x =	x_1	x_2	x_3	⋯	x_N	x_{N+1}	x_{N+2}	⋯	x_{N+M}
Labels	y =	-1	+1	-1		-1	?	?		?

Training data: each image patch is labeled as containing the object or background

Test data

- Classification function

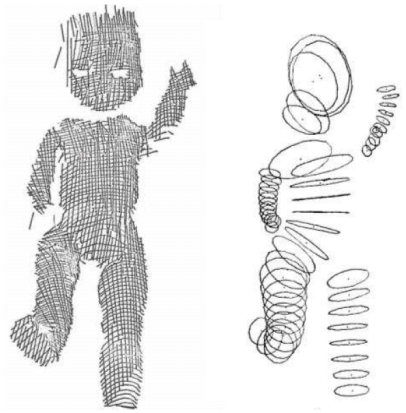
$$\hat{y} = F(x)$$ Where $F(x)$ belongs to some family of functions

- Minimize misclassification error

(Not that simple: we need some guarantees that there will be generalization)

Object representations

Explicit 3D models: use volumetric representation. Have an explicit model of the 3D geometry of the object.



Appealing but hard to get it to work…

Object representations

Implicit 3D models: matching the input 2D view to view-specific representations.

(b) For cars, classifiers are trained on 8 viewpoints

Not very appealing but somewhat easy to get it to work…

Class experiment

Class experiment

Experiment 1: draw a horse (the entire body, not just the head) in a white piece of paper.

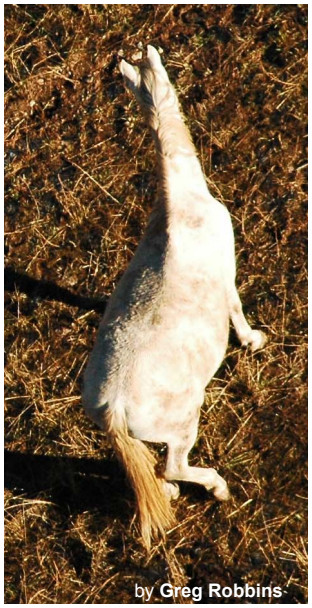
Do not look at your neighbor! You already know how a horse looks like… no need to cheat.

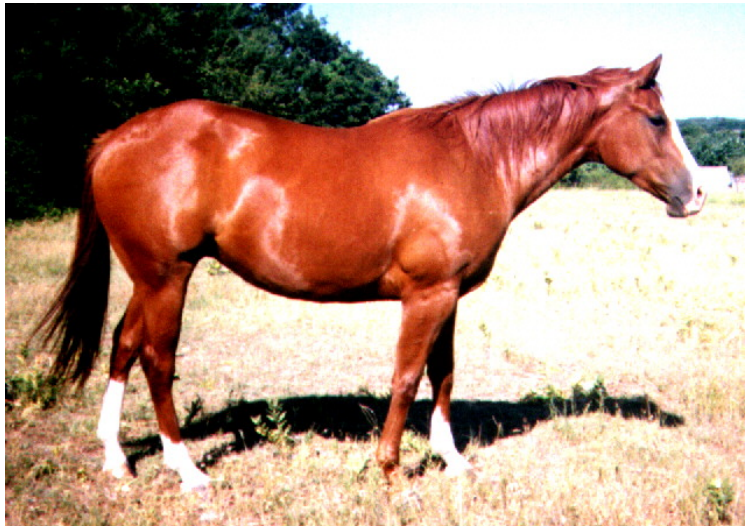
Class experiment

Experiment 2: draw a horse (the entire body, not just the head) but this time chose a viewpoint as weird as possible.

3D object categorization

Despite we can categorize all three pictures as being views of a horse, the three pictures do not look as being equally typical views of horses. And they do not seem to be recognizable with the same easiness.





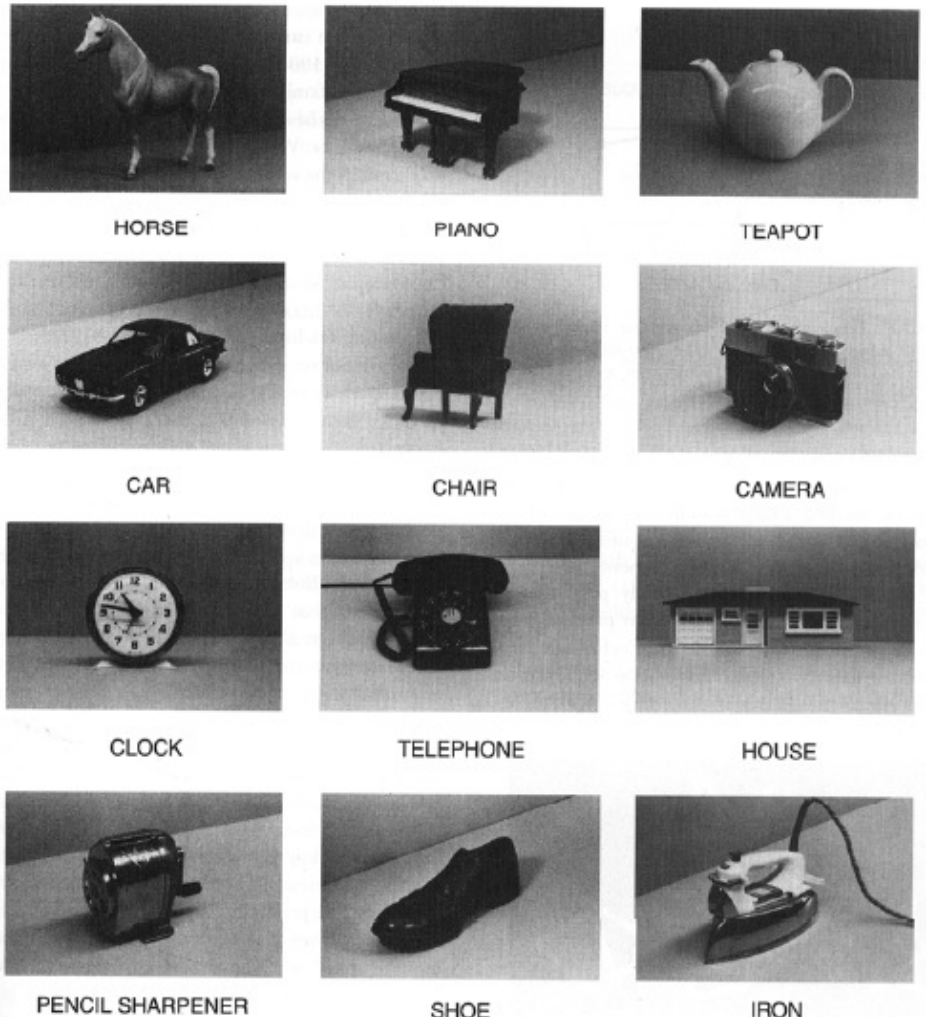
Canonical Perspective

Experiment (Palmer, Rosch & Chase 81): participants are shown views of an object and are asked to rate “how much each one looked like the objects they depict” (scale; 1=very much like, 7=very unlike)

In a recognition task, reaction time correlated with the ratings.

Canonical views are recognized faster at the entry level.

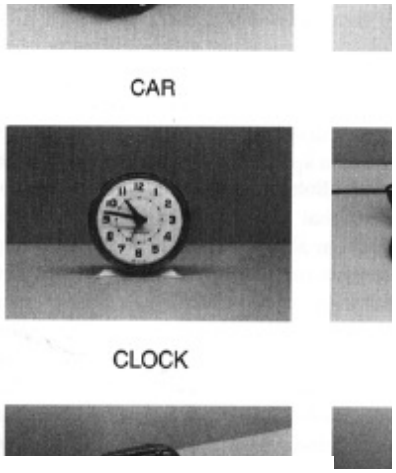
Examples of canonical perspective:

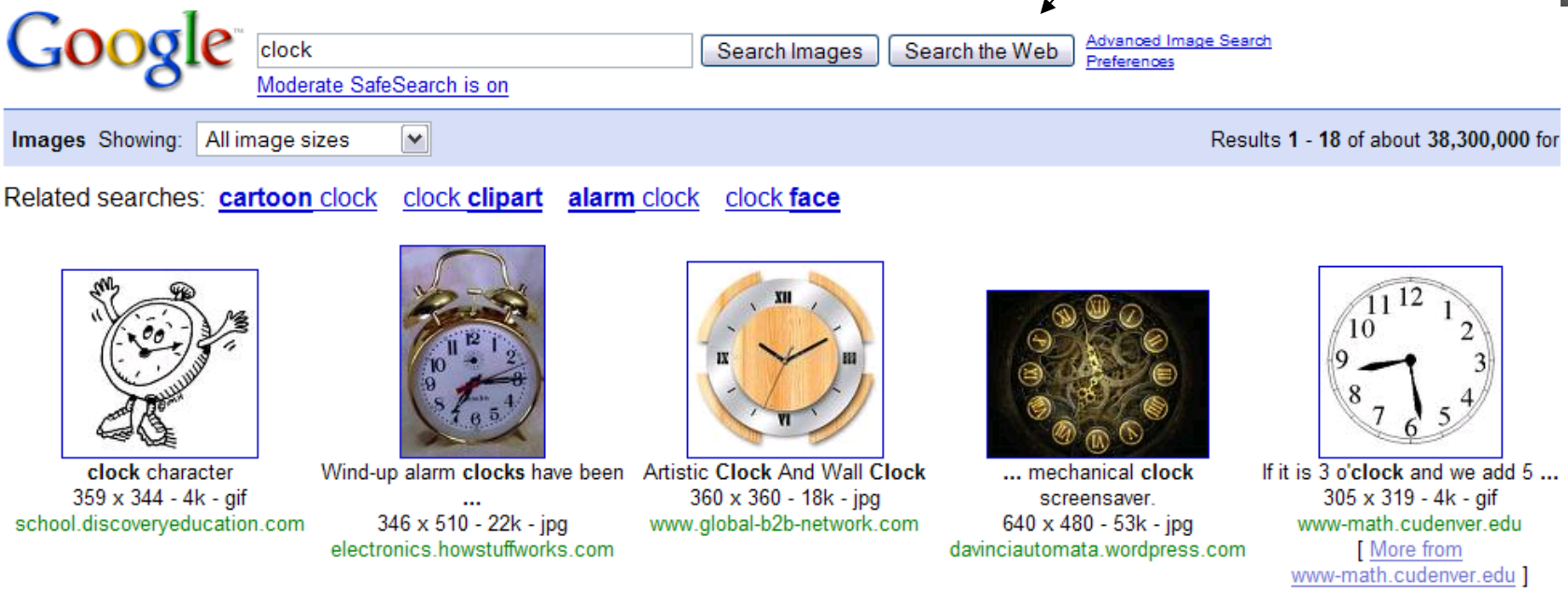


From *Vision Science*, Palmer

Canonical Viewpoint

Clocks are preferred as purely frontal





CVPR 2005

Histograms of Oriented Gradients for Human Detection

Navneet Dalal and Bill Triggs
INRIA Rhône-Alps, 655 avenue de l'Europe, Montbonnot 38334, France
{Navneet.Dalal,Bill.Triggs}@inrialpes.fr, http://lear.inrialpes.fr

Abstract

We study the question of feature sets for robust visual object recognition, adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

1 Introduction

Detecting humans in images is a challenging task owing to their variable appearance and the wide range of poses that they can adopt. The first need is a robust feature set that allows the human form to be discriminated cleanly, even in cluttered backgrounds under difficult illumination. We study

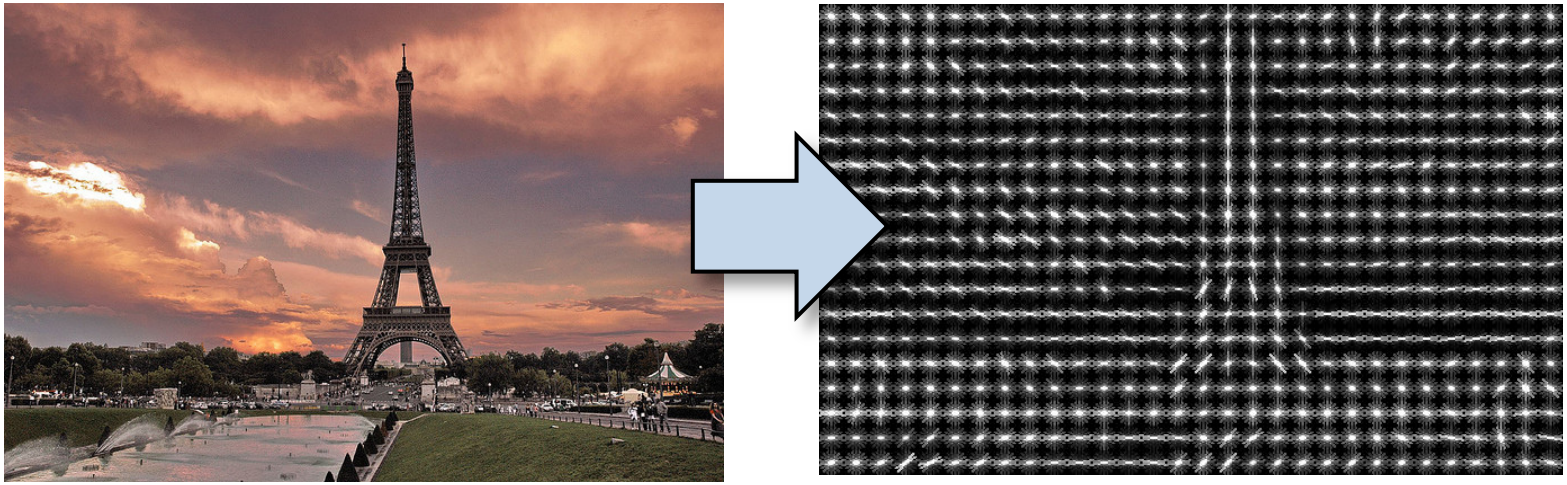
We briefly discuss previous work on human detection in §2, give an overview of our method §3, describe our data sets in §4 and give a detailed description and experimental evaluation of each stage of the process in §5–6. The main conclusions are summarized in §7.

2 Previous Work

There is an extensive literature on object detection, but here we mention just a few relevant papers on human detection [18,17,22,16,20]. See [6] for a survey. Papageorgiou *et al* [18] describe a pedestrian detector based on a polynomial SVM using rectified Haar wavelets as input descriptors, with a parts (subwindow) based variant in [17]. Depoortere *et al* give an optimized version of this [2]. Gavrila & Philomen [8] take a more direct approach, extracting edge images and matching them to a set of learned exemplars using chamfer distance. This has been used in a practical real-time pedestrian detection system [7]. Viola *et al* [22] build an efficient moving person detector, using AdaBoost to train a chain of progressively more complex region rejection rules based on Haar-like wavelets and space-time differences. Ronfard *et al* [19] build an articulated body detector by incorporating SVM based limb classifiers over 1st and 2nd order Gaussian

Human detection with HOG: Basic Steps

1. Map image to feature Space (HOG)



Human detection with HOG: Basic Steps

1. Map image to feature Space (HOG)
2. Training with positive and negative (linear SVM)

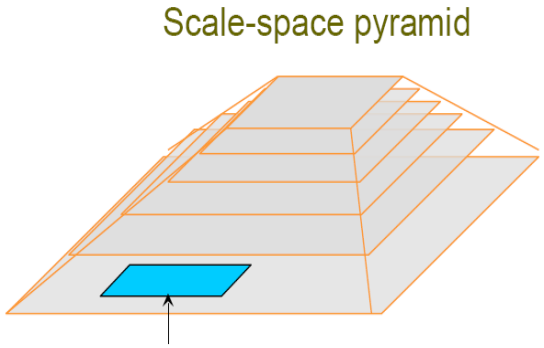
positive training examples

negative training examples

Human detection with HOG: Basic Steps

1. Map image to feature Space (HOG)
2. Training with positive and negative (linear SVM)
3. Testing : scan image in all scale and all location

   Binary classification on each location



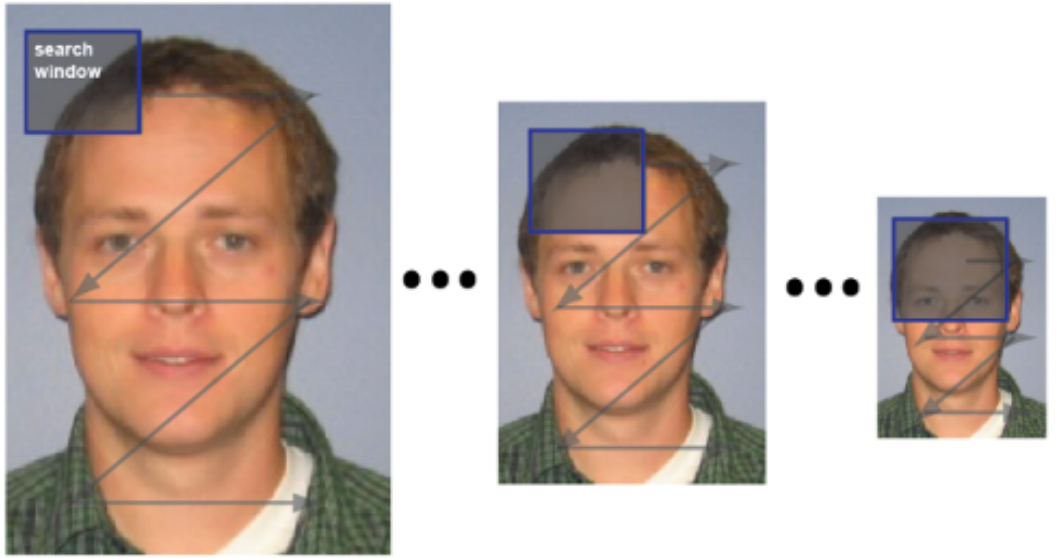


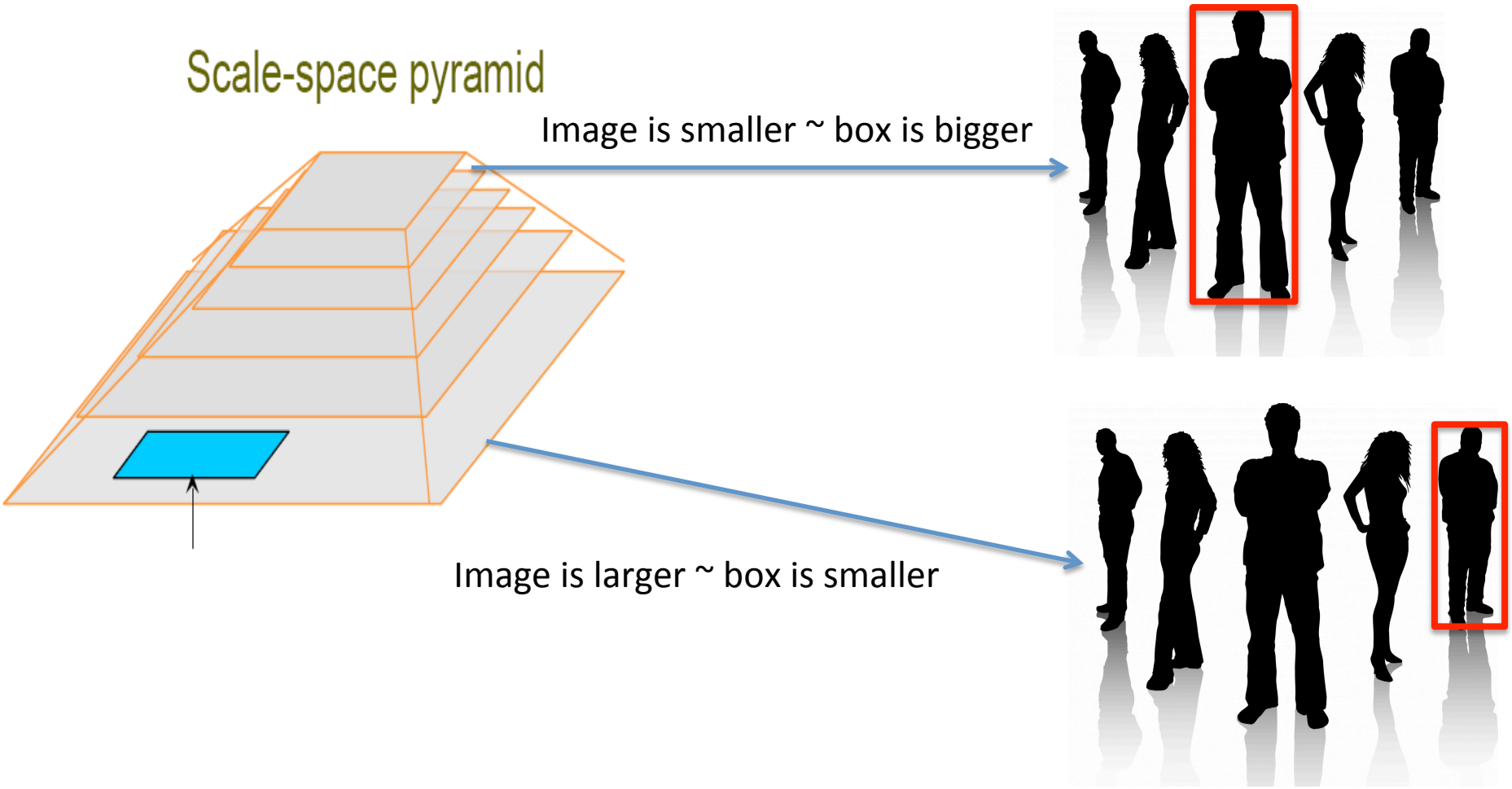
Image pyramid

Problem :
Bounding box size is different for the same object (different depth)

Solution 1:
Resize the box and do multiple convolution?
Not ideal :
It will change the feature dimension, need to retrain the SVM for each scale.

Image pyramid

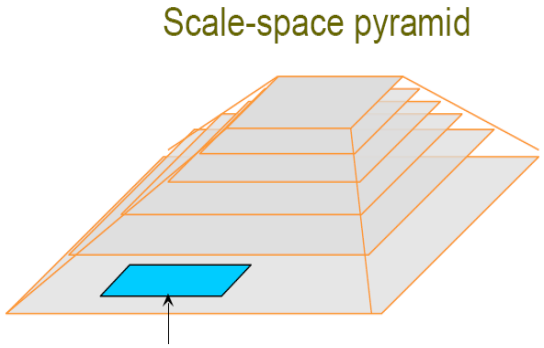
Solution 2:
Resize the image and do multiple convolution?–> image pyramid

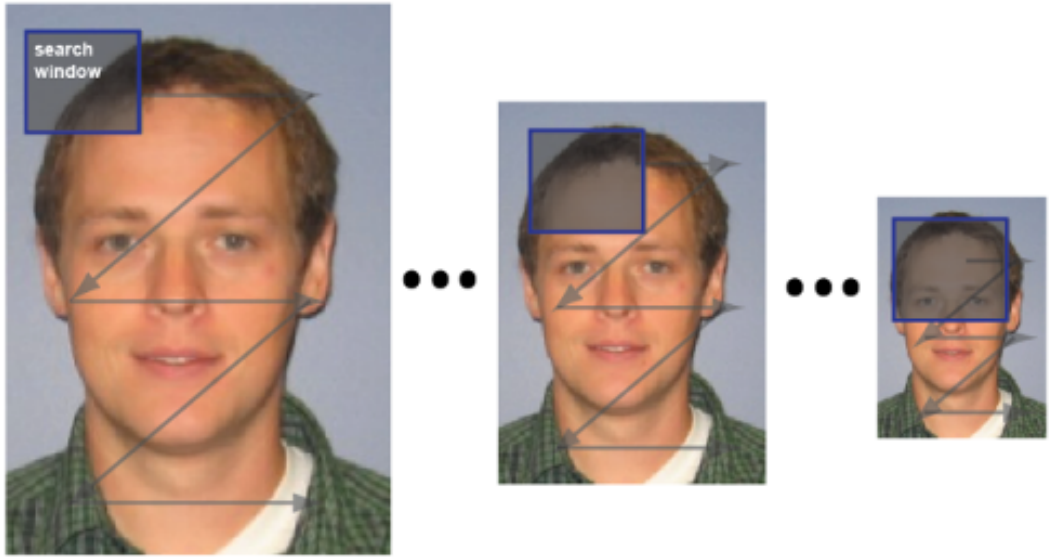


Human detection with HOG: Basic Steps

1. Map image to feature Space (HOG)
2. Training with positive and negative (linear SVM)
3. Testing : scan image in all scale and all location

   Binary classification on each location

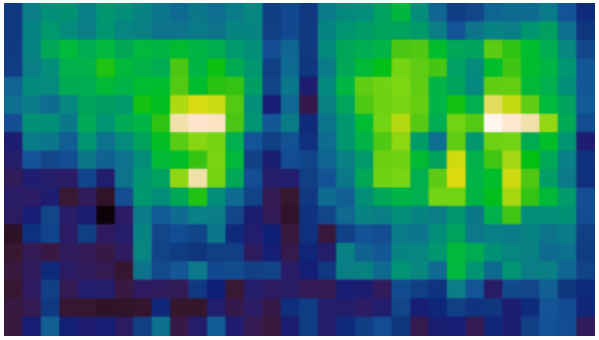


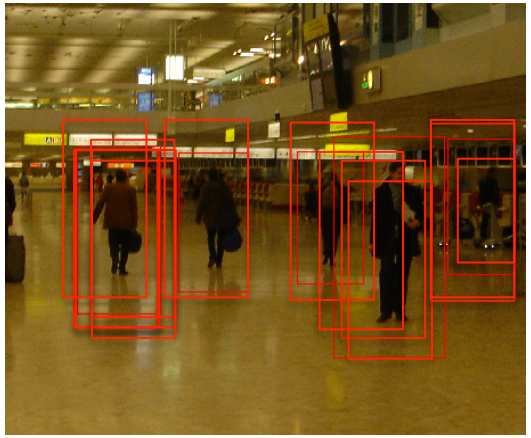


Human detection with HOG: Basic Steps

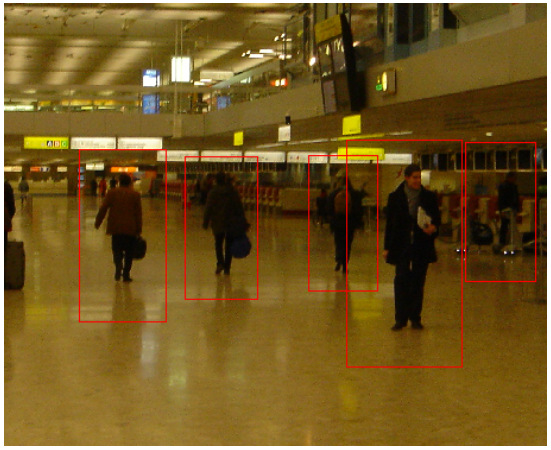
1. Map image to feature Space (HOG)
2. Training with positive and negative (linear SVM)
3. Testing : scan image in all scale and all location
4. Report box : non-maximum suppression

Detector response map





Final Boxes



After thresholding

After non-maximum suppression

Summary of Basic object detection Steps

Training:

Train a classifier describe the detection target

Testing :

Detection by binary classification on all location

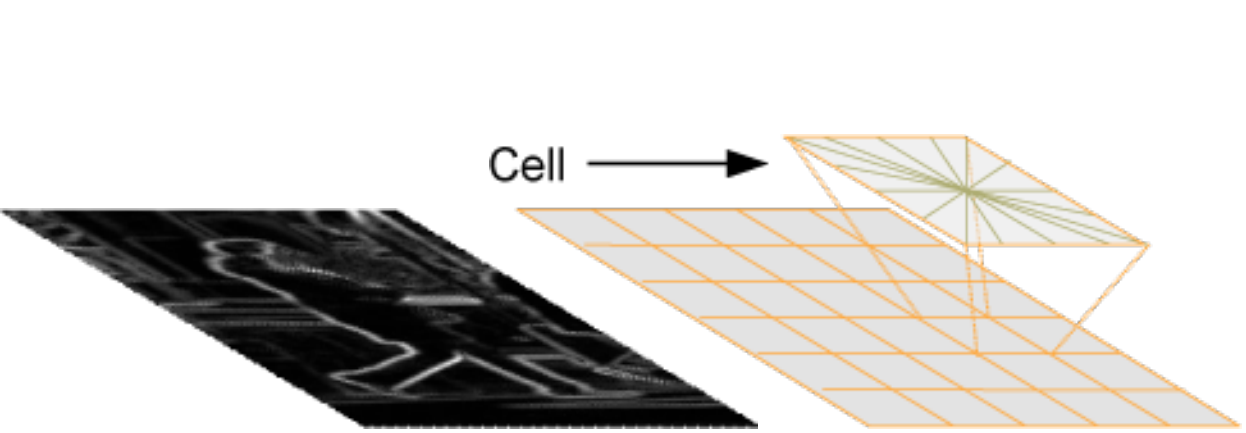
HOG descriptor

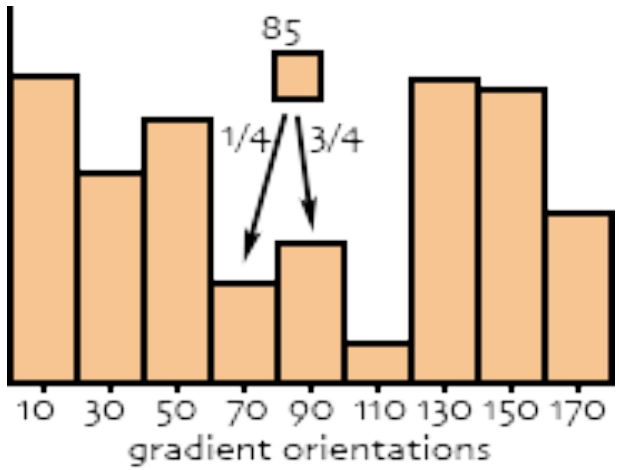
HOG： Gradients

- Compress image to 64x128 pixels
- Convolution with [-1 0 1] [-1 ; 0; 1] filters
- Compute gradient magnitude + direction
- For each pixel： take the color channel with greatest magnitude as final gradient

HOG: Cell histograms

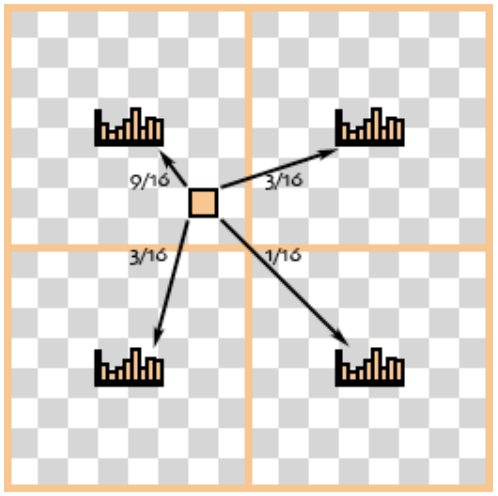
- Divide the image to cells, each cell 8x8 pixels
- Snap each pixel's direction to one of 18 gradient orientations
- Build histogram pre-cell using magnitudes

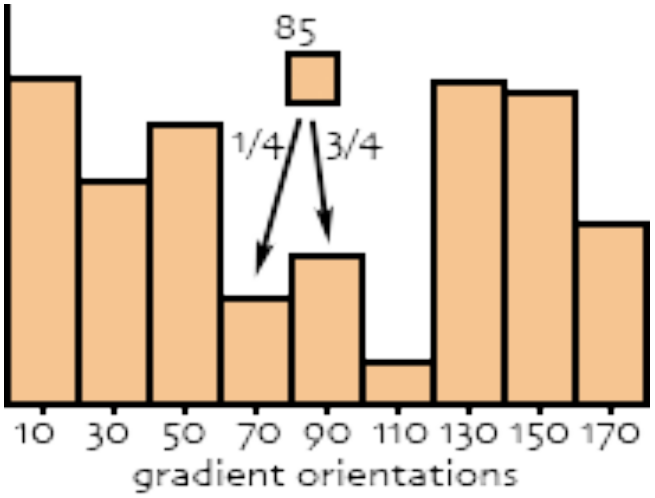




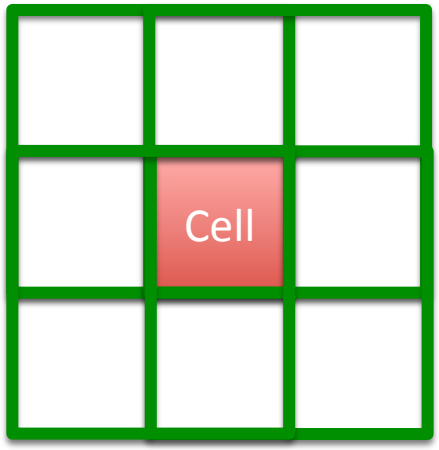
Histogram interpolation example

- Interpolated trilinearly:
  - Bilinearly into spatial cells
  - Linearly into orientation bins





Normalization



Current cell : 1x18 histogram

Block: 2x2 cell
overlapping with current cell

1. **contrast sensitive features:** 18 orientation -> 18 dim
2. **contrast insensitive features:** 9 orientation -> 9 dim

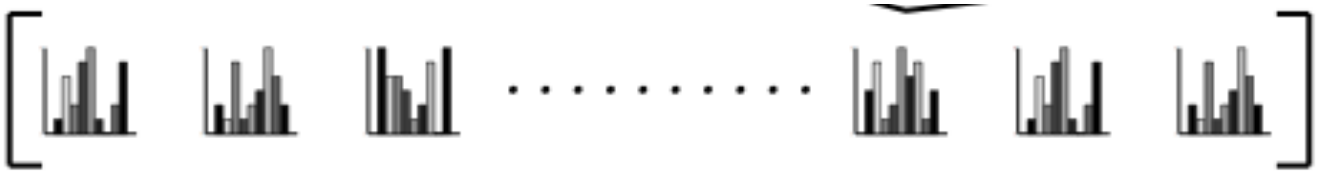
Normalize 4 times by its neighbor blocks, and average them

3**. texture features:** sum of the magnitude over all orientation and normalize 4 time, not average -> 4 dim

In total each cell : 18+9+4 dimension of feature

Final Descriptor

- Concatenation the normalized histogram

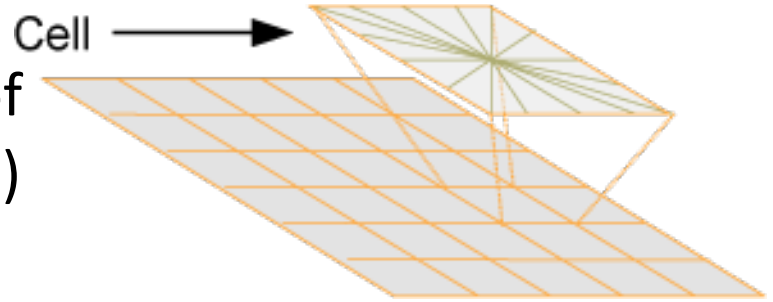


Visualization:

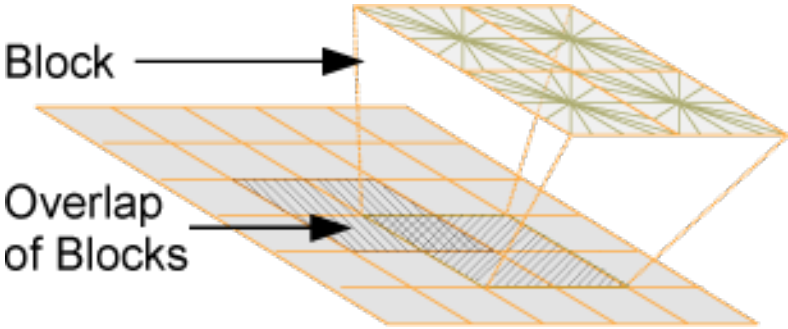
HOG Descriptor:

1. **Compute gradients** on an image region of 64x128 pixels

2. **Compute histograms** on 'cells' of typically 8x8 pixels (i.e. 8x16 cells)



3. **Normalize histograms** within overlapping blocks of cells



4. **Concatenate histograms**

It is a typical procedure of feature extraction !

Feature Engineering

- Developing a feature descriptor requires a lot of engineering
  - Testing of parameters (e.g. size of cells, blocks, number of cells in a block, size of overlap)
  - Normalization schemes
- An extensive evaluation was performed to make these design desiccations
- It's not only the idea, but also the engineering effort

Problem ?

Single, rigid template usually not enough to represent a category.

- Many object categories look very different from different viewpoints, or style

- Many objects (e.g. humans) are articulated, or have parts that can vary in configuration

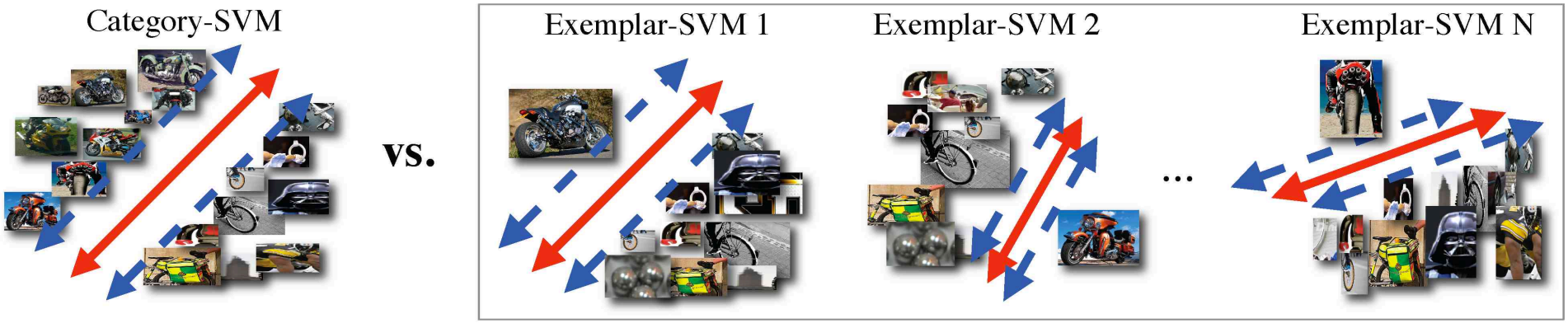
Solution :

- Exemplar SVM: Ensemble of Exemplar-SVMs for Object Detection and Beyond

- Part Based Model

Exemplar–SVM

- Still a rigid template, but train a separate SVM for each positive instance



For each category it can has exemplar with different size aspect ratio

Benefit from Exemplar–SVM ?

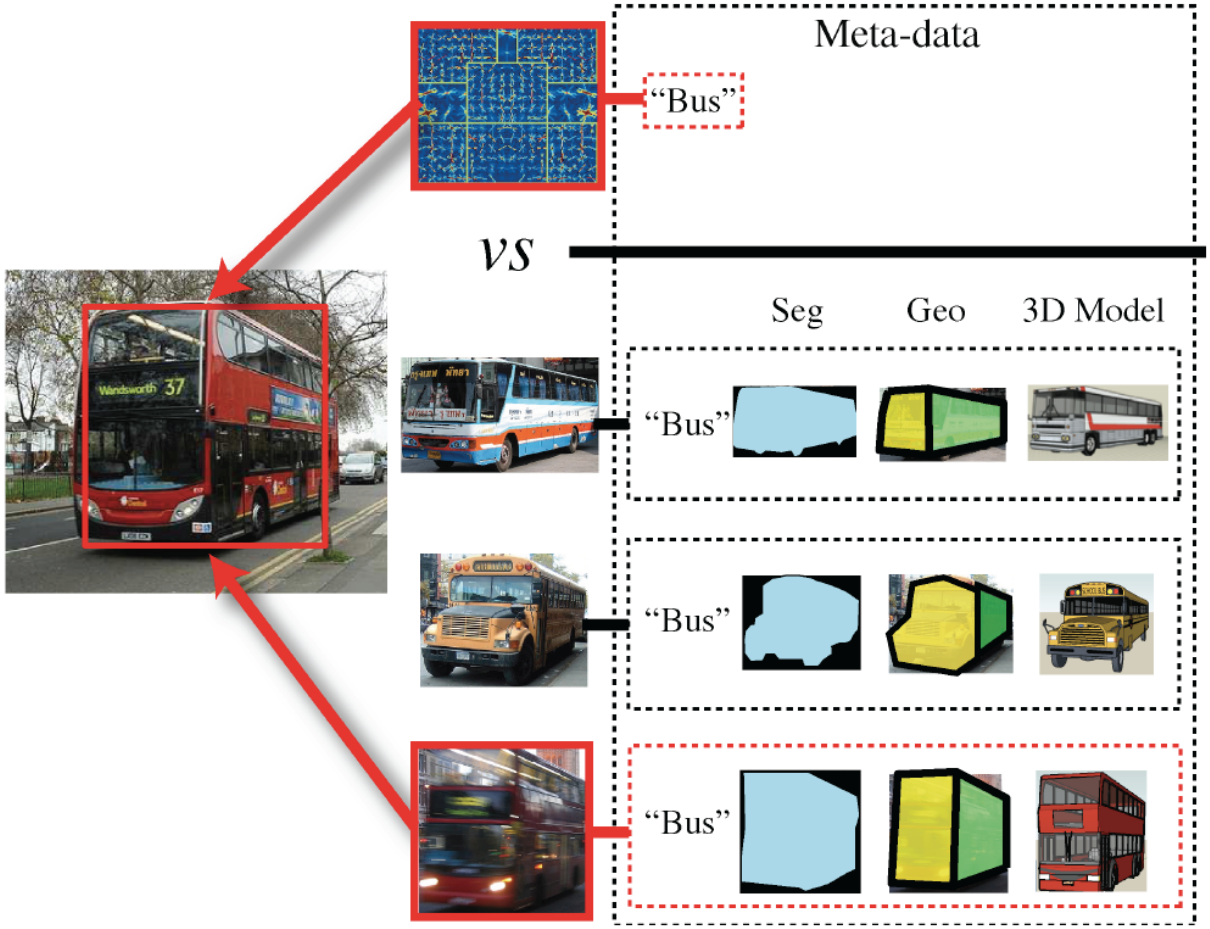
- Handel the intra-category variance naturally, without using complicated model.
- Compare to nearest neighbor approach : make use of negative data and train a discriminative object detector
- Explicit correspondence from detection result to training exemplar

Benefit from Exemplar–SVM ?

- Explicit correspondence from detection result to training exemplar

We not only know it is train, but also its orientation and type!

Benefit from Exemplar-SVM ?



Category-based Detector

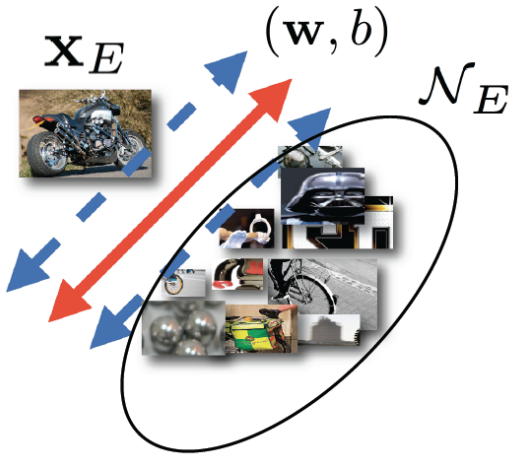
Ensemble of Exemplar-SVMs

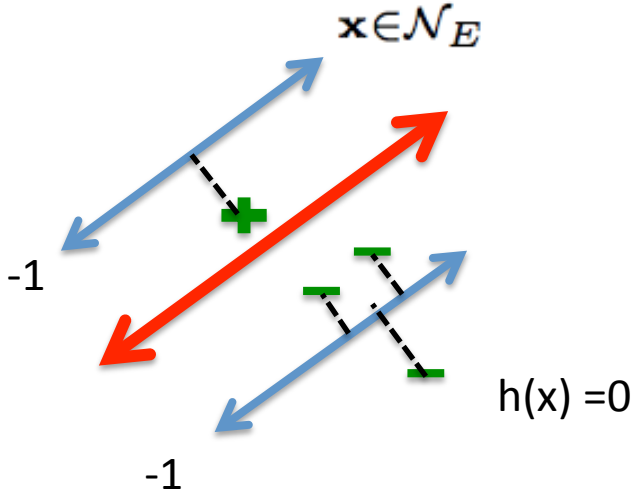
We can do even more

Training Exemplar–SVM

Objective Function:

$$\Omega_E(\mathbf{w}, b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T \mathbf{x}_E + b) + C_2 \sum_{\mathbf{x} \in \mathcal{N}_E} h(-\mathbf{w}^T \mathbf{x} - b)$$





Learn the w that minimize the objective function, equivalent to maximize the margin

Hard Negative Mining

$$\Omega_E(\mathbf{w}, b) = ||\mathbf{w}||^2 + C_1 h(\mathbf{w}^T \mathbf{x}_E + b) + C_2 \sum_{\mathbf{x} \in \mathcal{N}_E} h(-\mathbf{w}^T \mathbf{x} - b)$$

$\mathcal{N}_E$

Windows from images not containing any in-class instances: but there is too many!
2,000 images x 10,000 windows per image = 20M negatives

Find ones that you get wrong by a search, and train on these hard ones

Hard Negative Mining

Input : Positive : exemplar E

Negative : images and bounding boxes for this category

$N=\{(J_1,B_1), (J_2,B_2),\ldots(J_m,B_m)\}$

Initialize : random pick m patches N_{random} from N that not overlap with

[SV,b,w]=trainSVM(E, N_{random})

Hard negative mining

While : i != m or N_{hard} not empty

for i= 1to n do

D = detect(b,w,J_i)

N_i= D.conf > threshold & D not overlap with B_i

Add Ni to N_{hard}

if |N_{hard}| > memory−limit, **then** break;

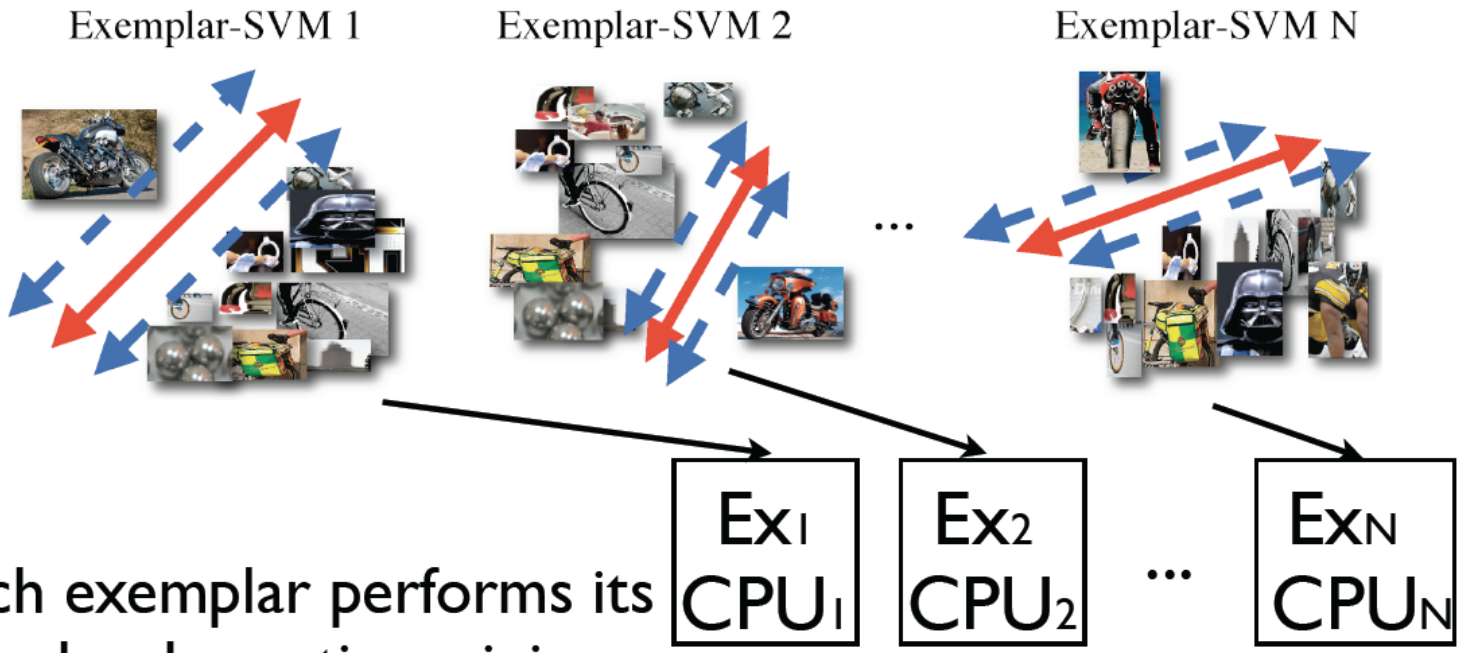
end

[SV^{new},b^{new},w^{new}]=trainSVM(E, [$N_{random,}$SV])

SV = [SV; SV^{new}]

end

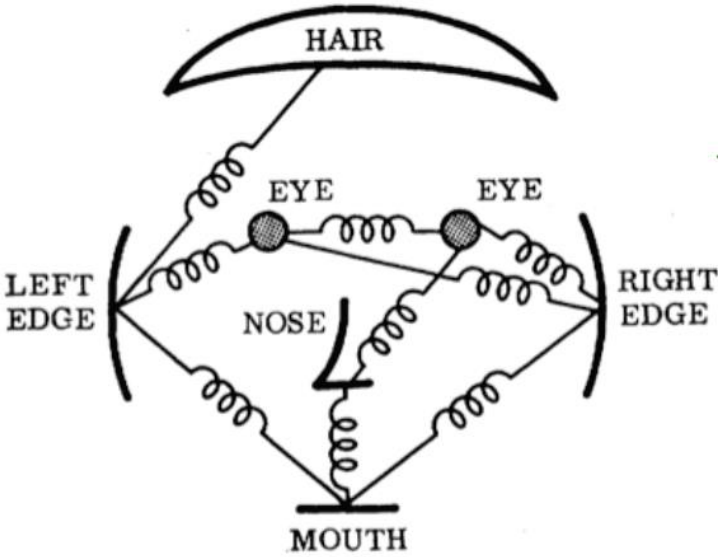
Embarrassingly Parallel



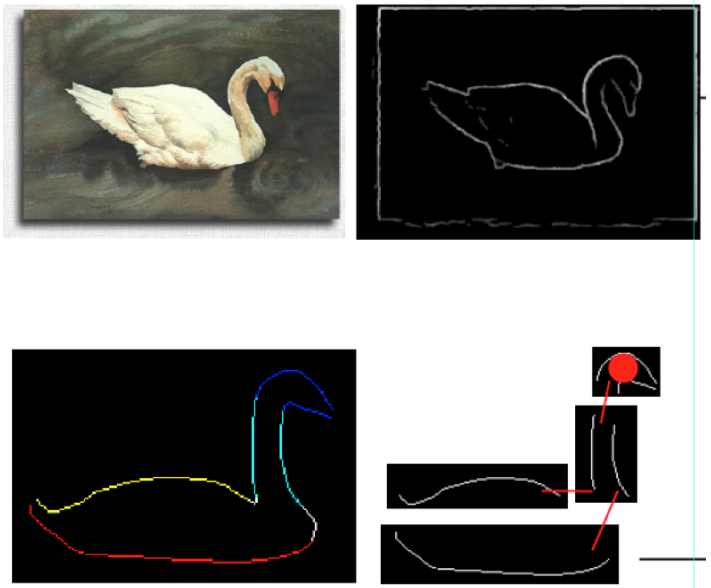
- Each exemplar performs its own hard negative mining
- Solve many convex learning problems
- Parallel training on cluster

Part Based detector

Objects are represented by features of parts and spatial relations between parts



Face model by Fischler and Elschlager ‘73

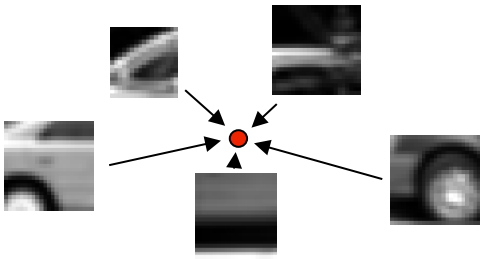


Part Based detector

- How to defined the parts for one object category
- How to represent their spatial relation shape
- How to combine parts detection and spatial relations to obtained the final detection

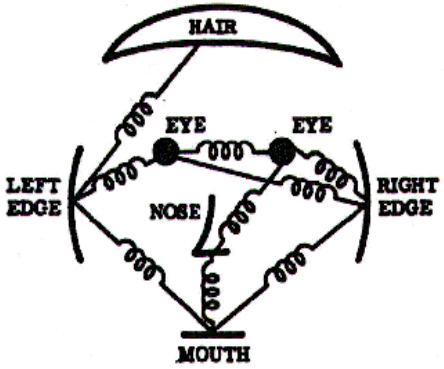
Structure models

Voting models



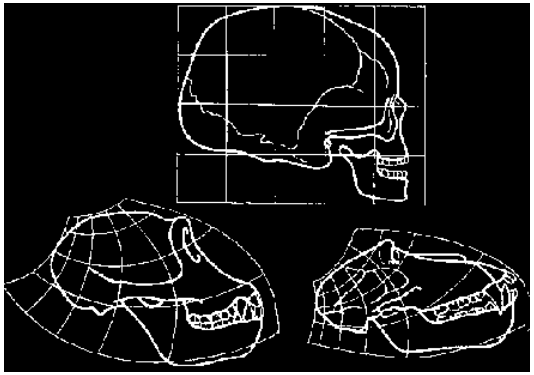
• Many parts (>100)

Constellation models

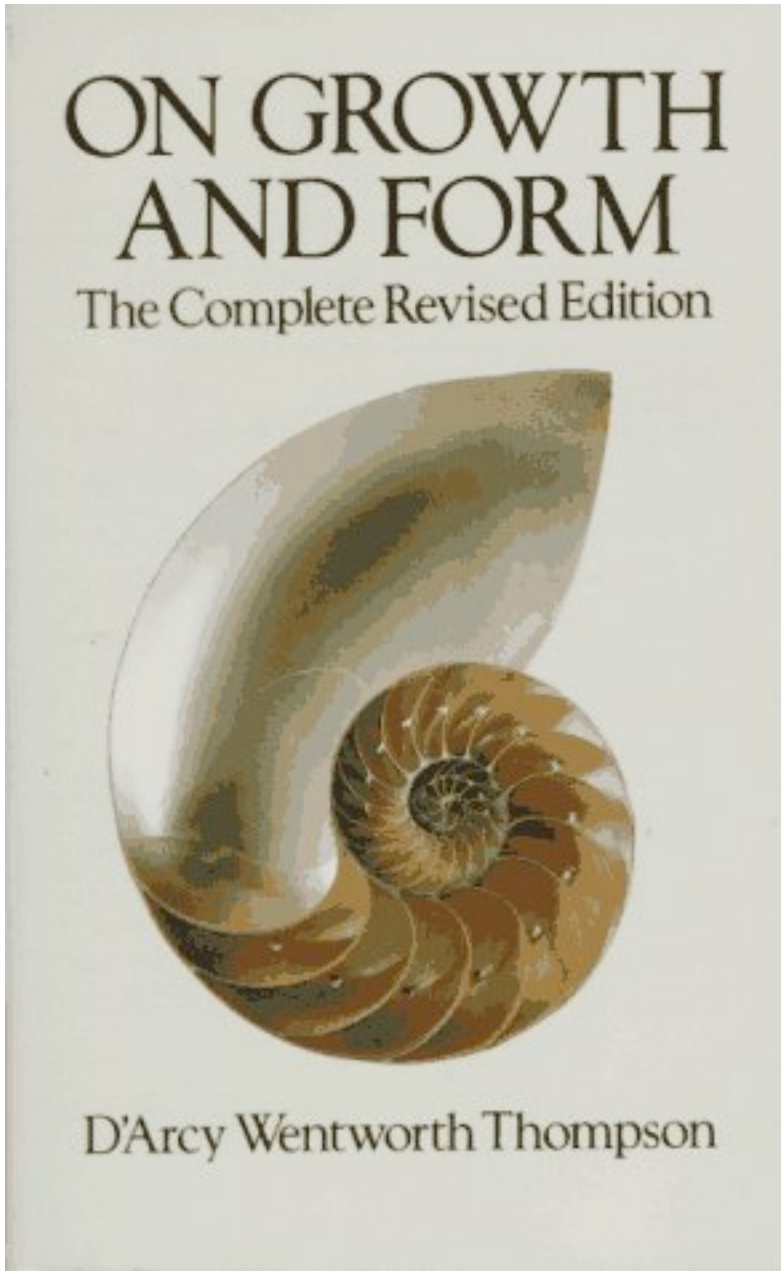


• Few parts (~6)

Deformable models



• No parts



to the lines of our new curved ordinates. In like manner, the still more bizarre outlines of other fishes of the same family of Chaetodonts will be found to correspond to very slight modifications of

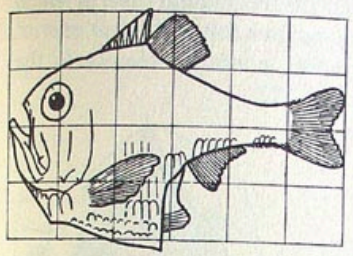


Fig. 146. *Argyropelecus olfersi.*

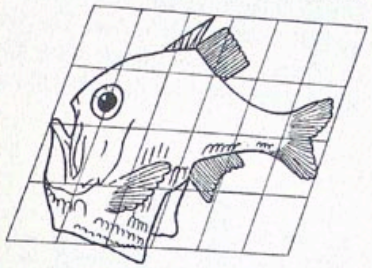


Fig. 147. *Sternoptyx diaphana.*

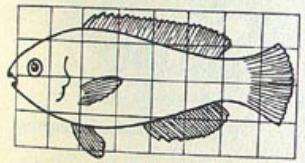


Fig. 148. *Scarus* sp.

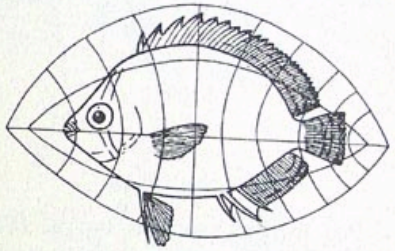


Fig. 149. *Pomacanthus.*

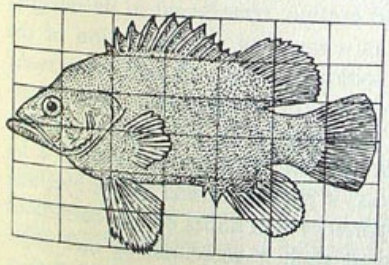


Fig. 150. *Polyprion.*

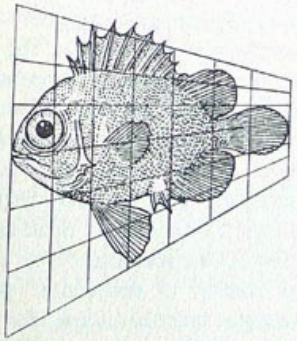


Fig. 151. *Pseudopriacanthus altus.*

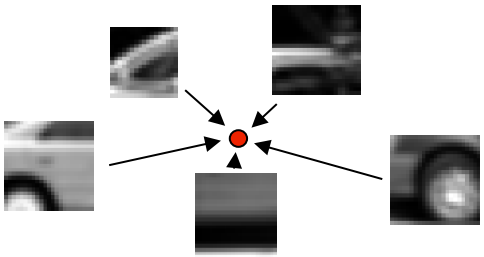
similar co-ordinates; in other words, to small variations in the values of the constants of the coaxial curves.

In Figs. 150–153 I have represented another series of Acanthopterygian fishes, not very distantly related to the foregoing. If we

From wikipedia: Perhaps the most famous part of the work is chapter XVII, "The Comparison of Related Forms," where Thompson explored the degree to which differences in the forms of related animals could be described by means of relatively simple mathematical transformations.

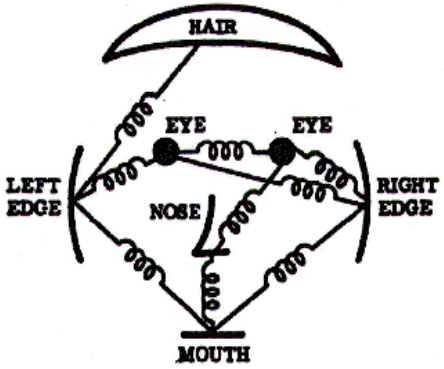
Structure models

Voting models



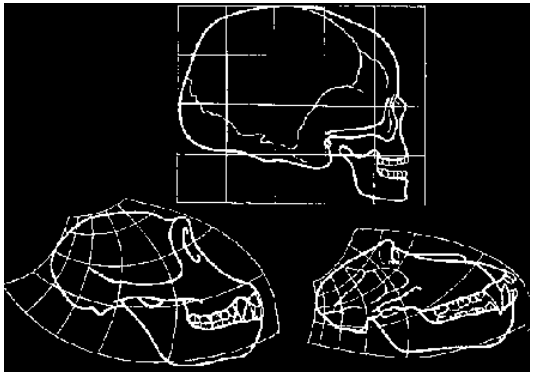
• Many parts (>100)

Constellation models



• Few parts (~6)

Deformable models



• No parts

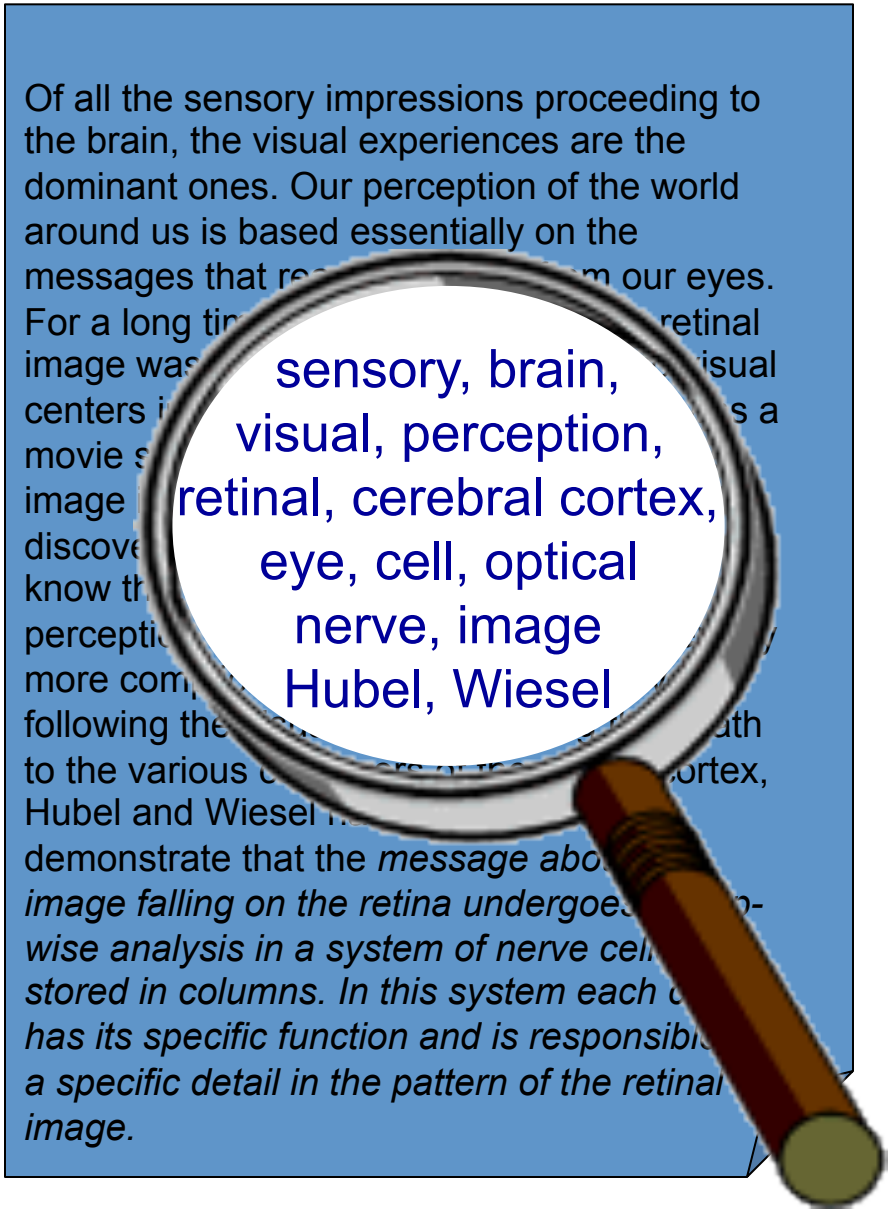
Bag of words

• No structure!

Object → Bag of ‘words’

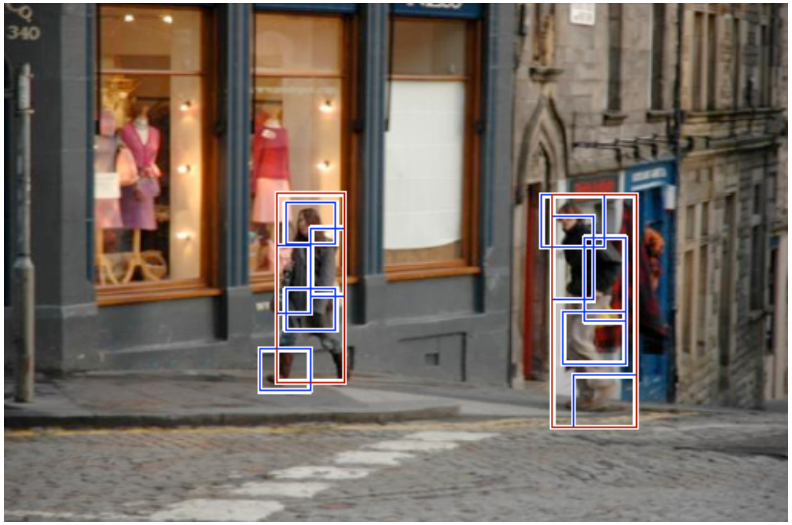
Slide credit: Fei fei

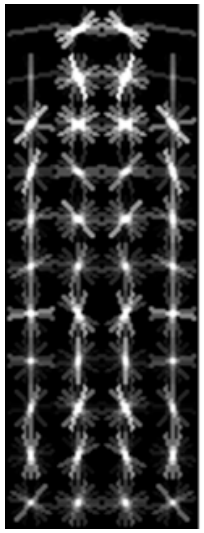
Analogy to documents

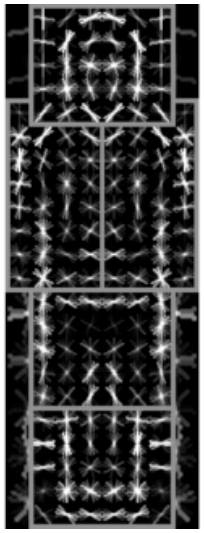


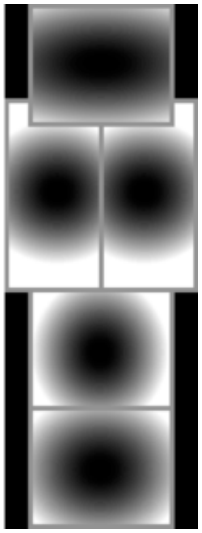
Slide credit: Fei fei

DPM : Object Detection with Discriminatively Trained Part Based Models









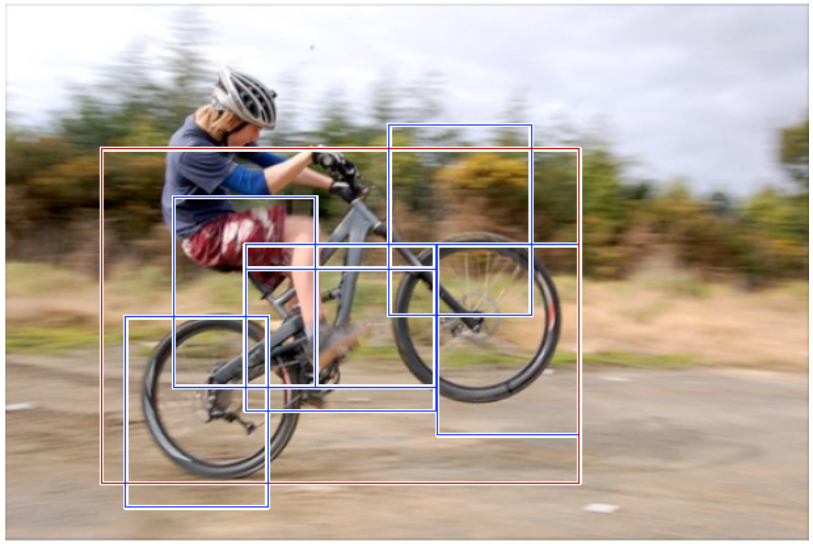
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan,
Object Detection with Discriminatively Trained Part Based Models, PAMI 32(9), 2010

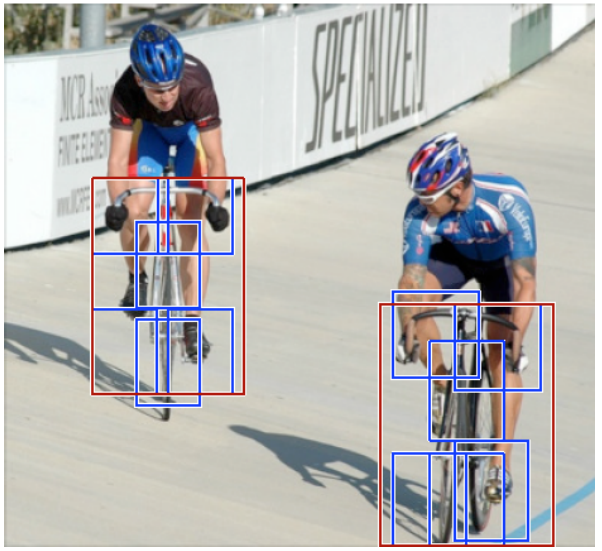
DPM: overview

- Each category detector has mixture of deformable part models (components)
- Each component has global template + deformable parts
- Fully trained from bounding boxes alone

(Latent SVM)

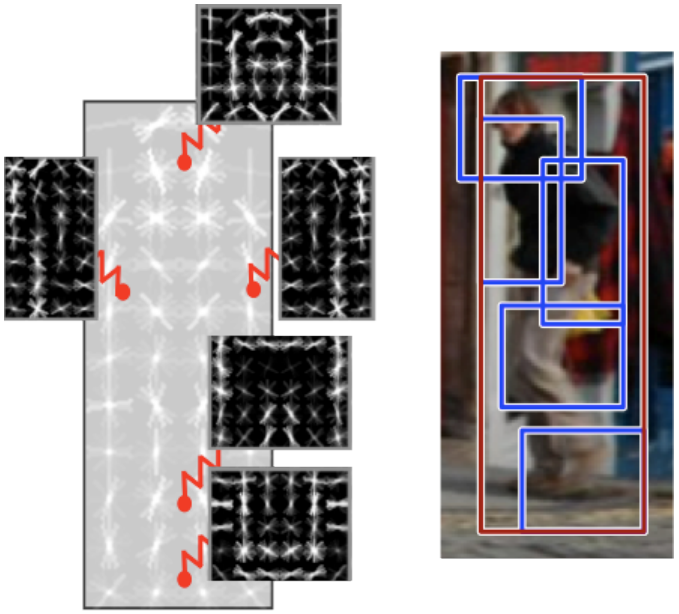
DPM: component





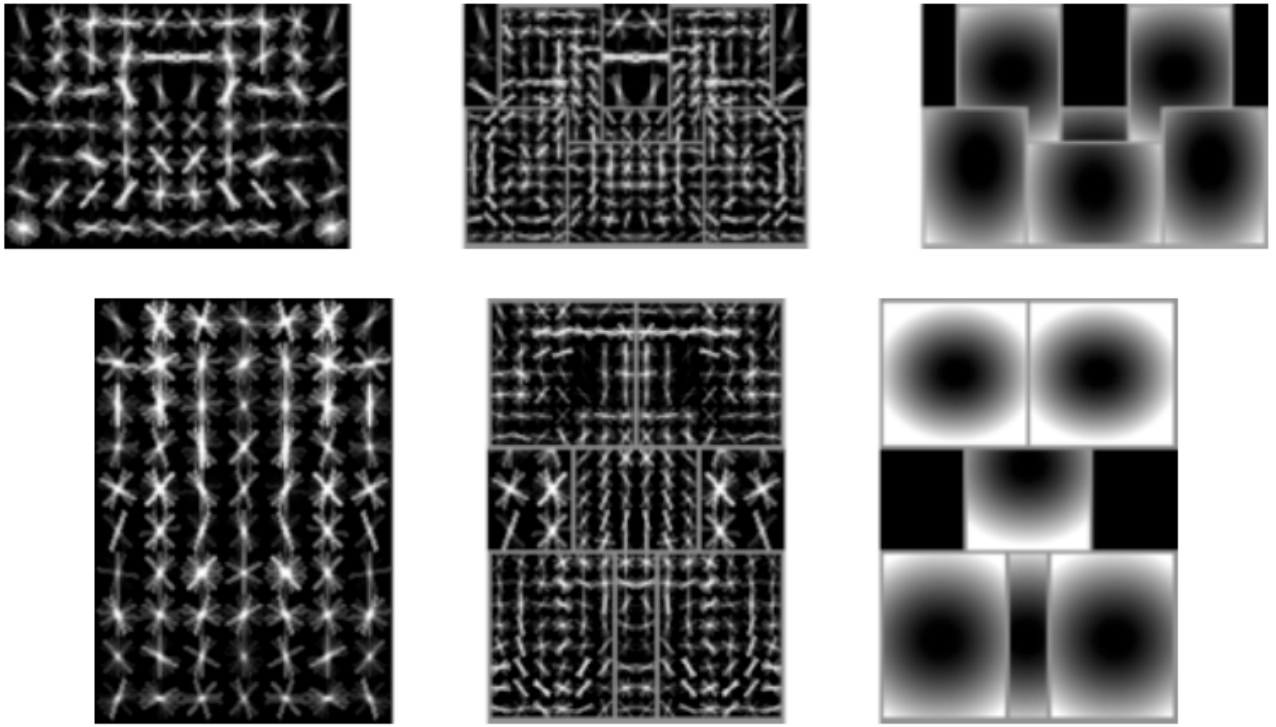
- Each category detector has mixture of component for different aspect ratio (handle intra-class variance)
- Each component has a it's own DPM model

Deformable part models



Model encodes local appearance + pairwise geometry

DPM: component



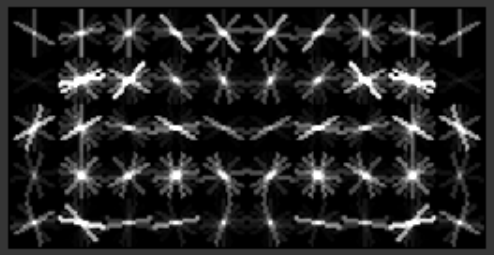
root filters
coarse resolution

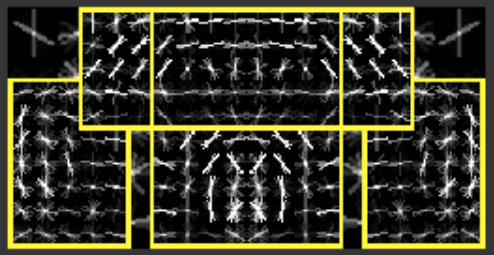
part filters
finer resolution

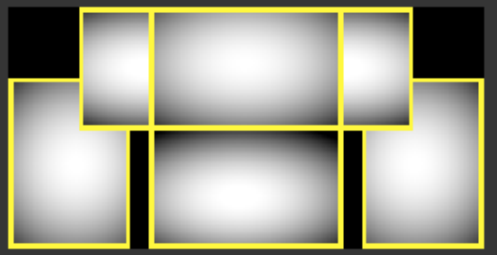
deformation
models

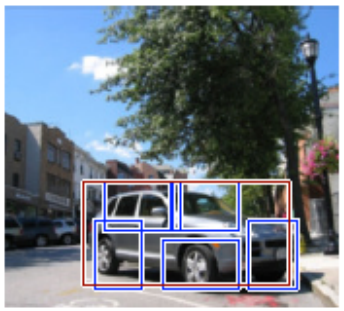
Each component has a root filter F_0
and n part models (F_i, v_i, d_i)

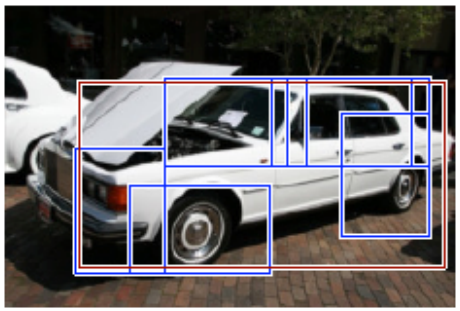
Example models

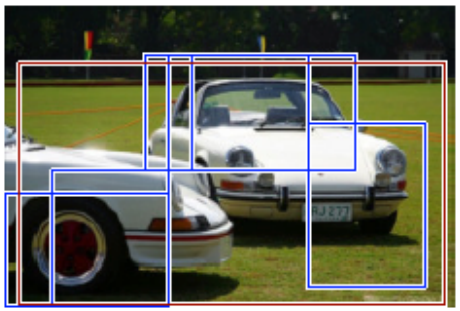


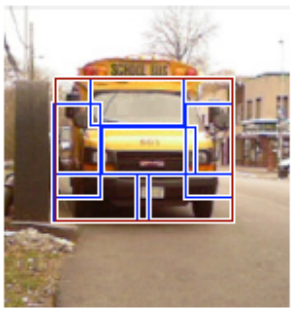






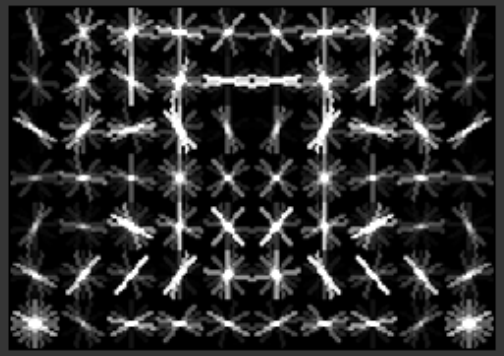


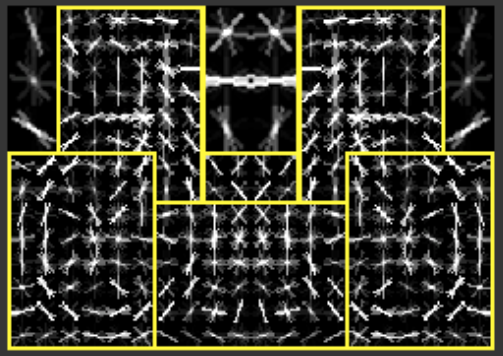


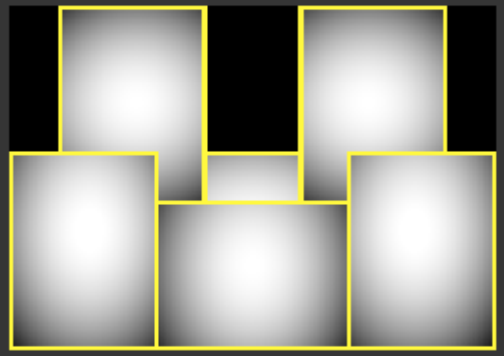


Source: Deva Ramanan

Example models

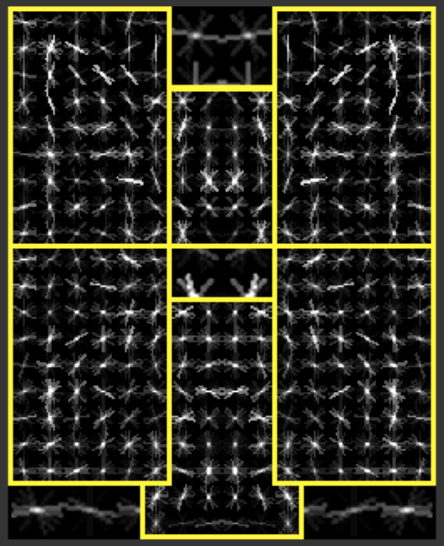


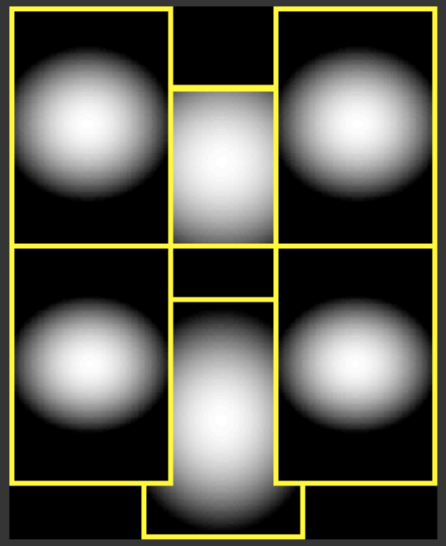


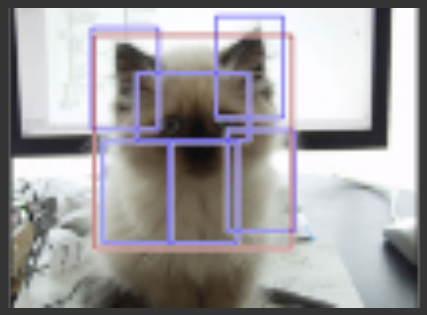


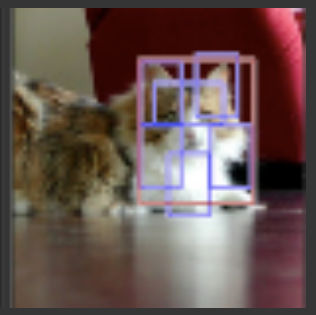
Source: Deva Ramanan

Example models



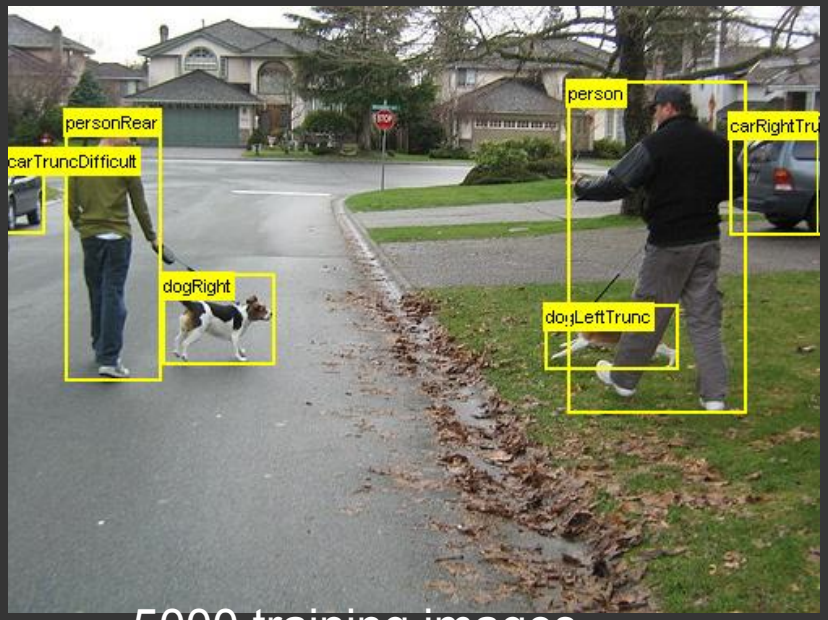




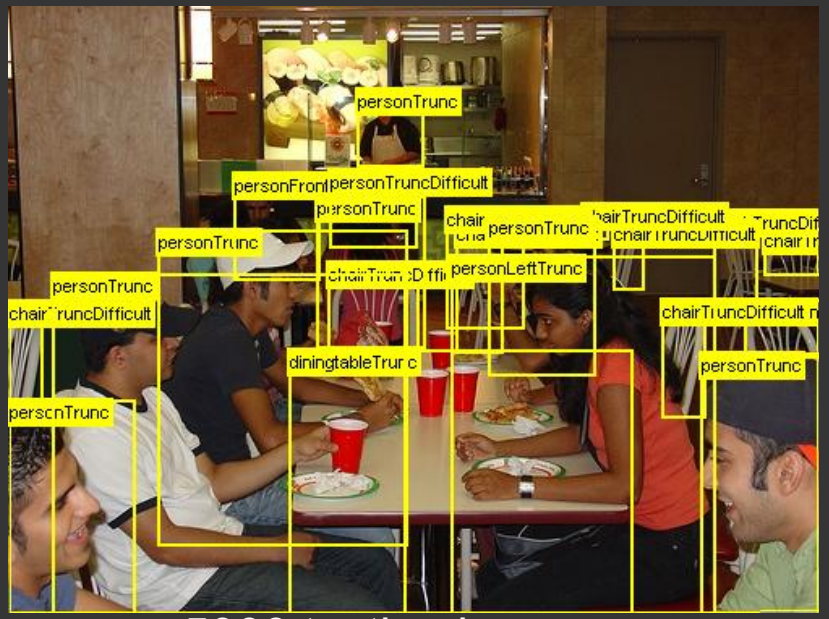



horse
sofa
bottle
cat
AUG V INIA 07
FIGHT TERRORISM

PASCAL Visual Object Challenge



5000 training images



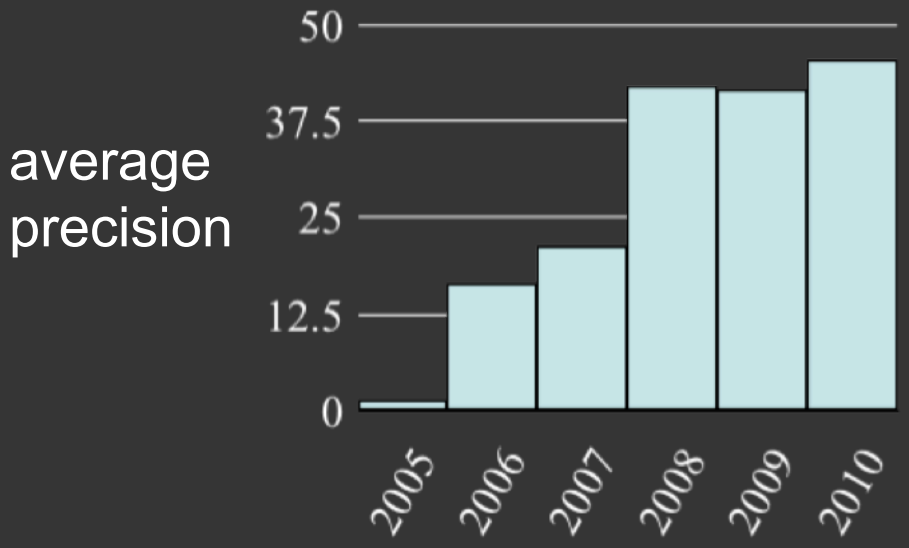
5000 testing images

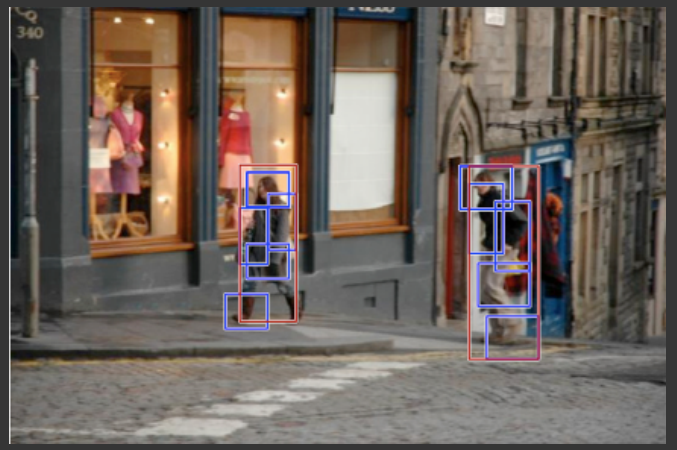
20 everyday object categories

aeroplane bike bird boat bottle bus car cat chair cow table
dog horse motorbike person plant sheep sofa train tv

Source: Deva Ramanan

5 years of PASCAL people detection





1% to 45% in 5 years

Discriminative mixtures of star models 2007-2010 Felzenszwalb, McAllester, Ramanan CVPR 2008
Felzenszwalb, Girshick, McAllester, and Ramanan PAMI 2009

Source: Deva Ramanan

Evaluation: Precision & Recall

	Patient has the disease	Patient doesn't have disease
Test is positive	Correct result	False positive
Test is negative	False negative	Correct result

Truth

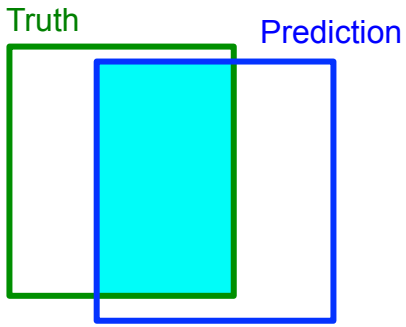
Object No Object

Algorithm Prediction → Object

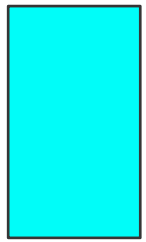
Algorithm Prediction → No Object

Object	True Positive	False Positive
No Object	False Negative	True Negative

Precision = TP / (TP + FP) Recall = TP / (TP + FN)

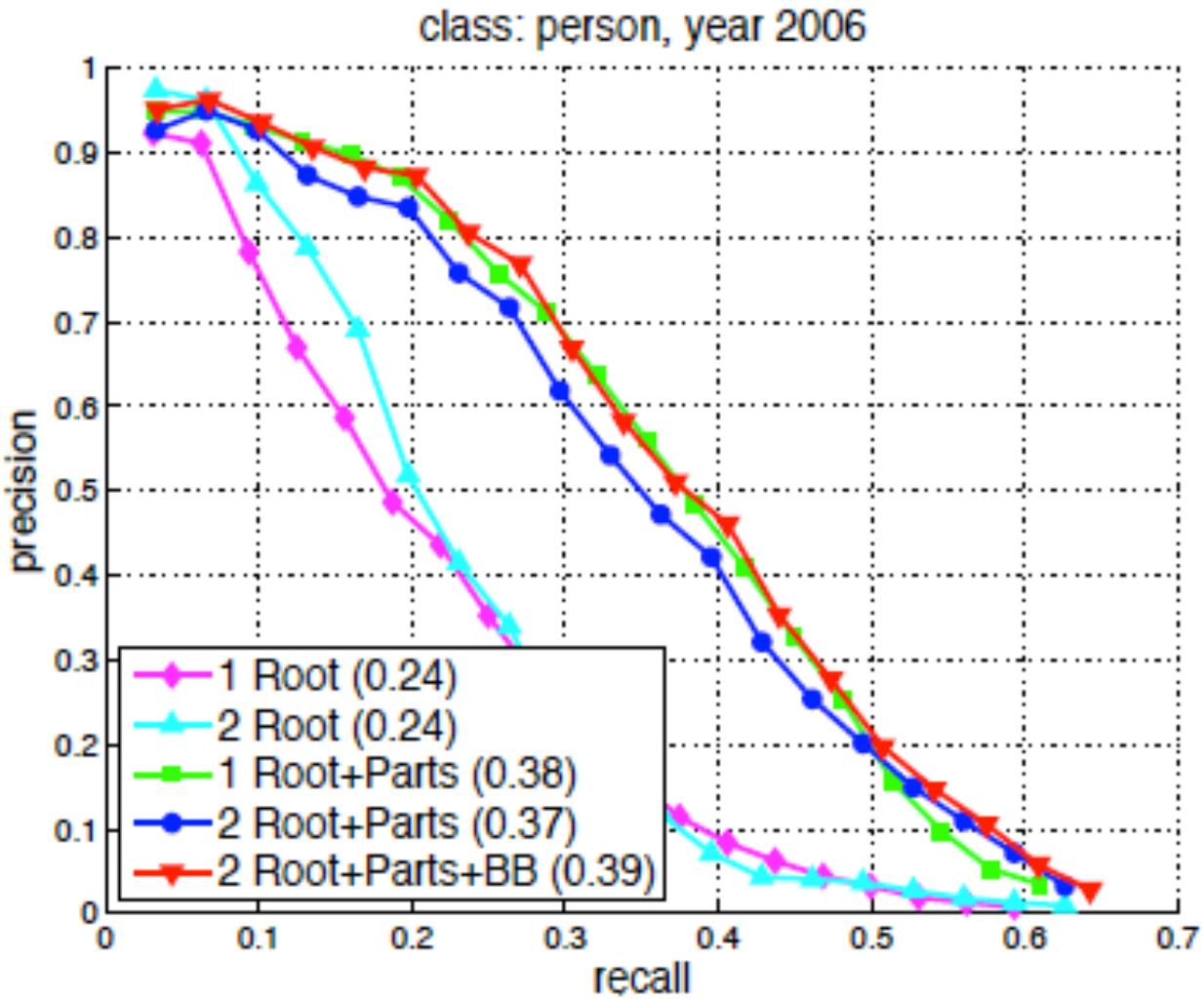


Intersection Over Union =



/

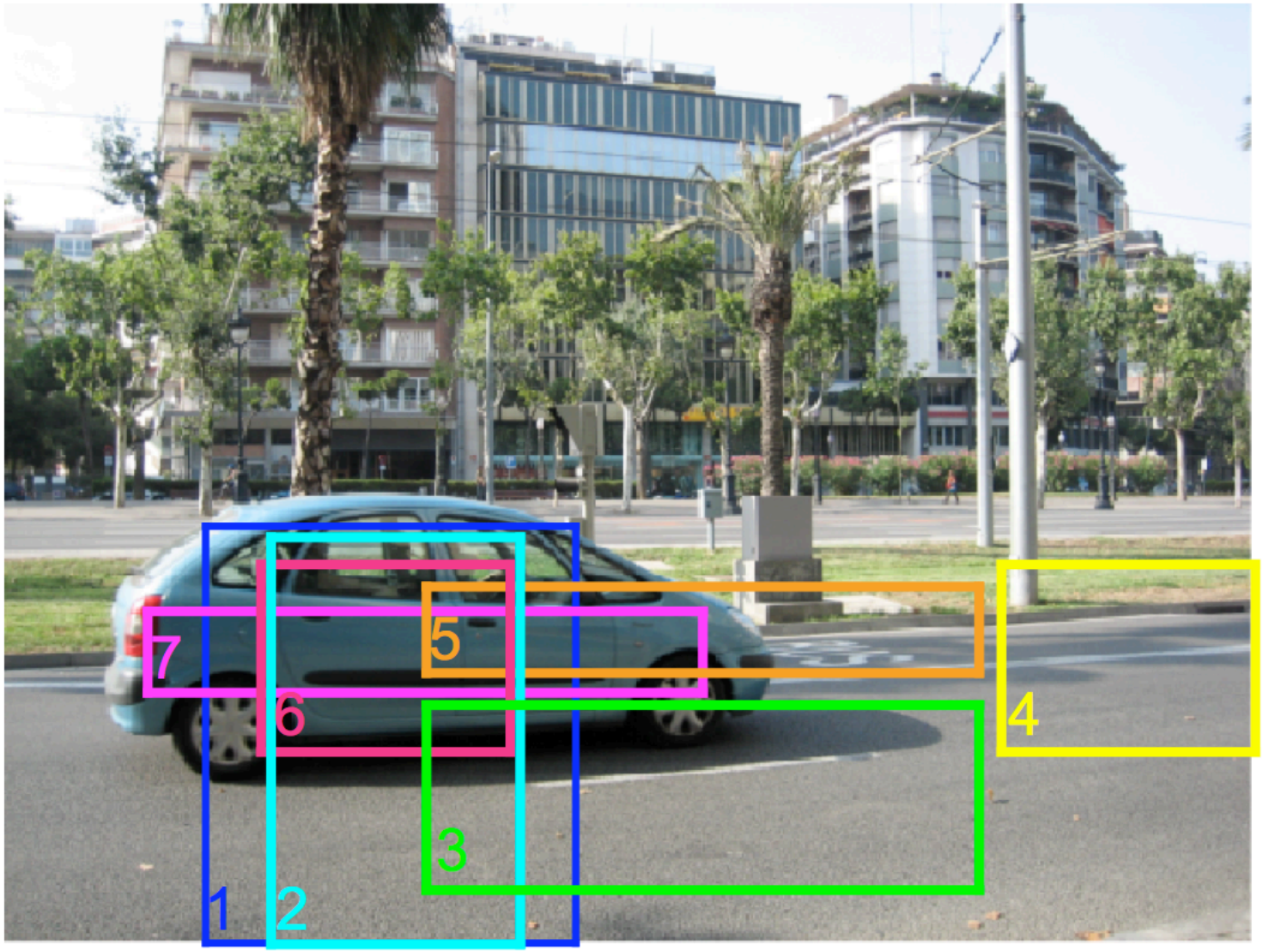
Evaluation: Precision & Recall



Evaluation of performance

Before plotting and ROC or precision-recall curves…

The detector challenge: by looking at the output of a detector on a random set of images, can you guess which object is it trying to detect?



1. chair, 2. table, 3. road, 4. road, 5. table, 6. car, 7. keyboard.

Concept Review

- Object vs. Scene vs. Texture
- Instances vs. categories
- Mediated vs. Directed
- Entry-level vs. Fine-grained
- Knowledge-based vs. Data-driven
- Explicit 3D vs. Implicit 3D (view-based)
- Whole vs. parts
- Discriminative vs. generative
- Structure vs. Bag-of-Words