Lecture 7
Introduction to Object Recognition

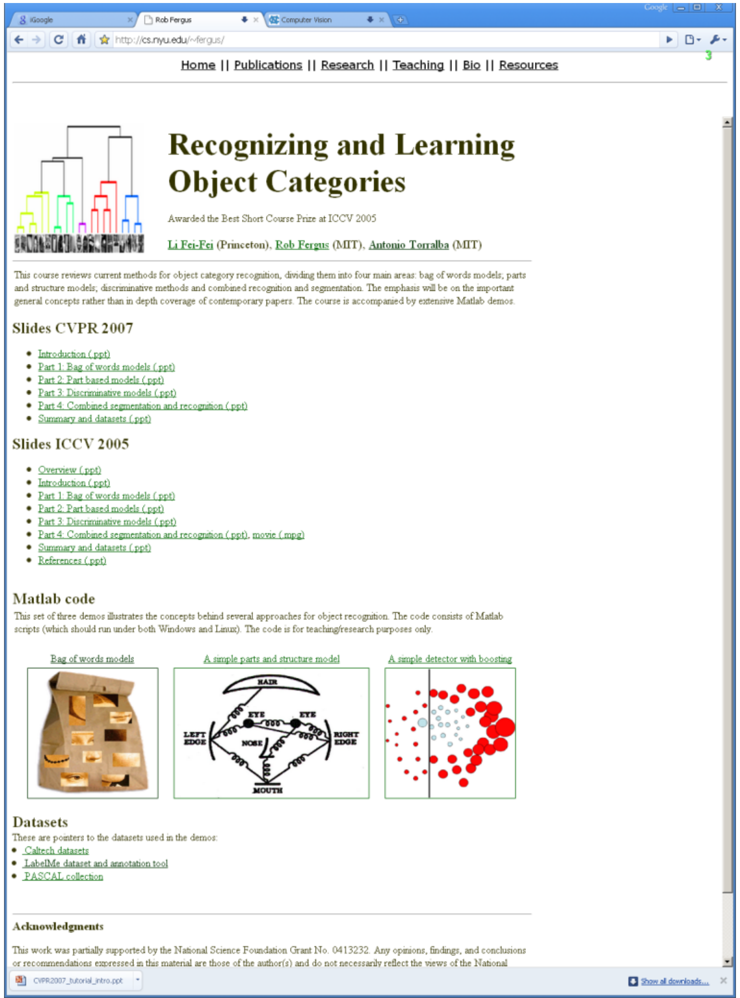
Slides from CVPR 2007 short course with Fei-Fei Li and Antonio Torralba; and also from Svetlana Lazebnik

Admin

- Assignment 2 is out.

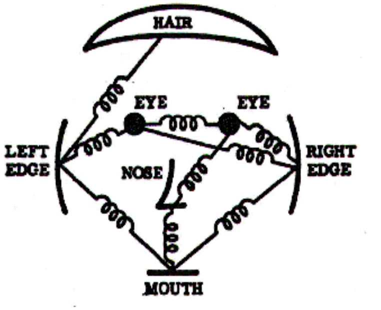
Short Course Webpage

http://people.csail.mit.edu/torralba/shortCourseRLOC

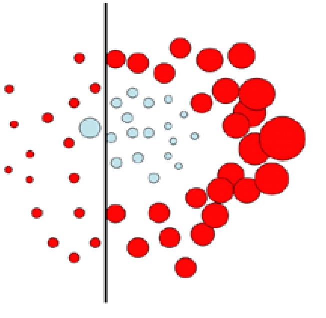


Agenda

- Introduction
- Bag-of-words models



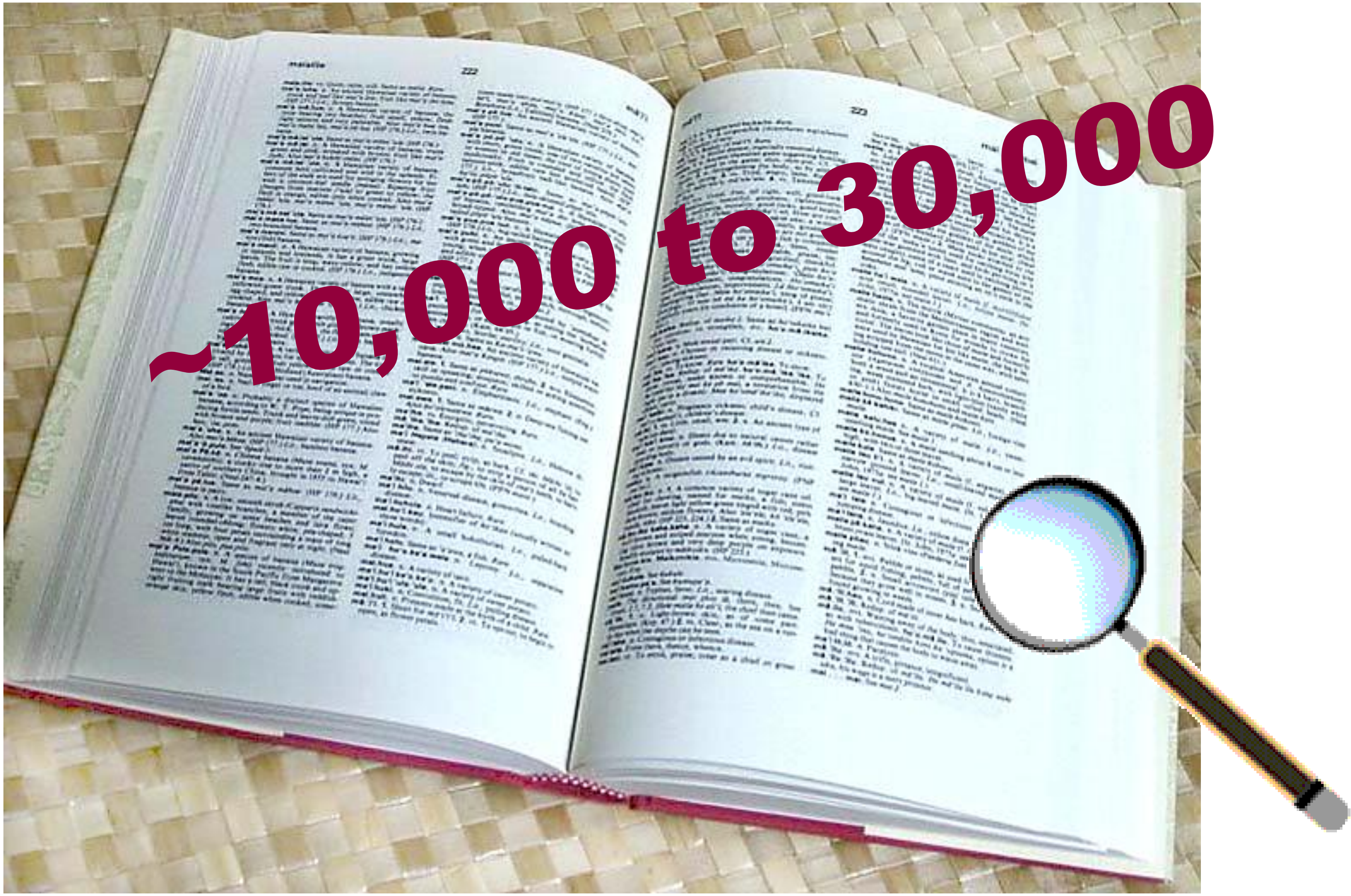
- Part-based models
- Discriminative methods



- Segmentation and recognition
- Datasets & Conclusions

Bruegel, 1564

How many object categories are there?



Biederman 1987

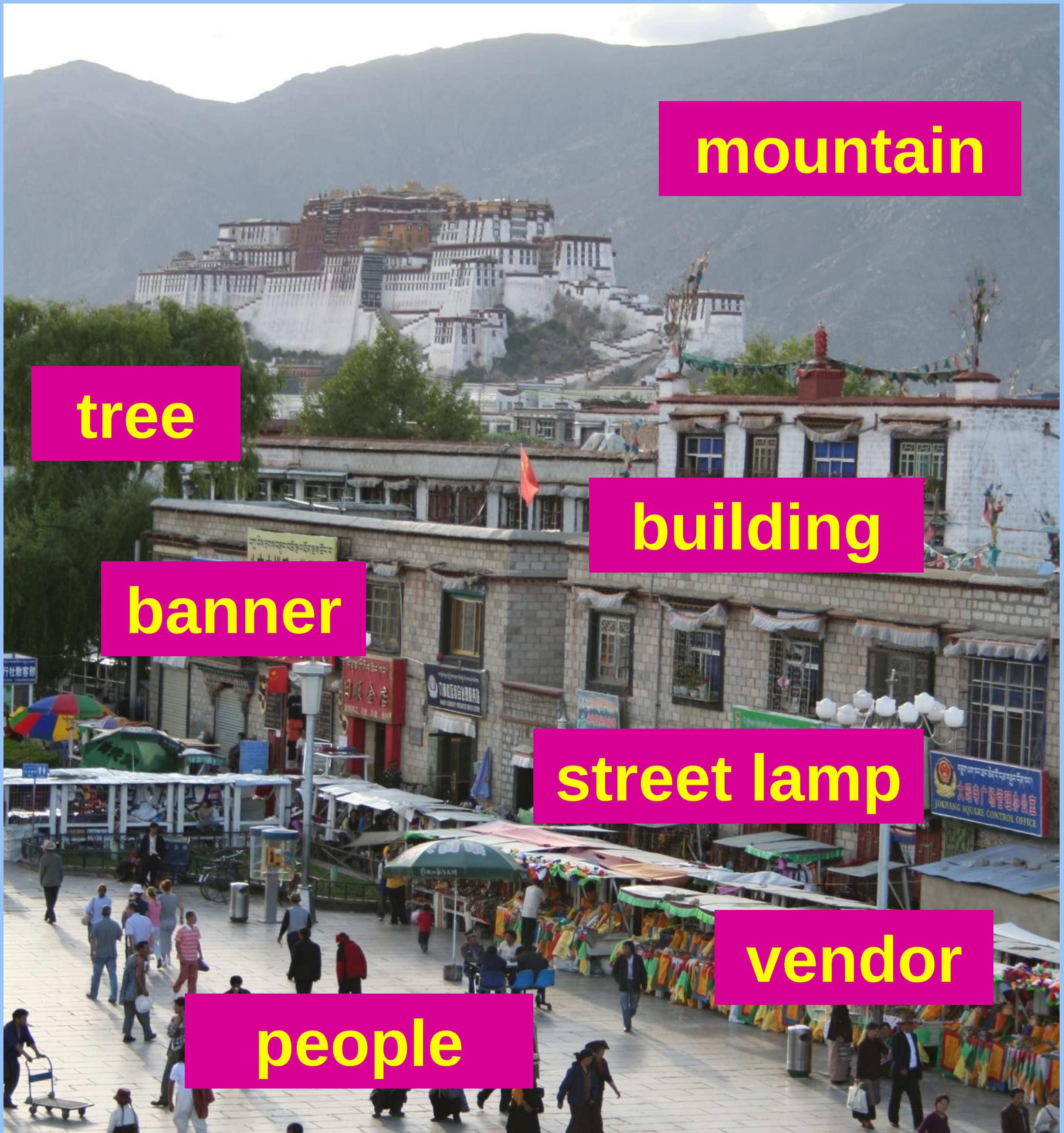
So what does object recognition involve?

Classification: does this contain people?

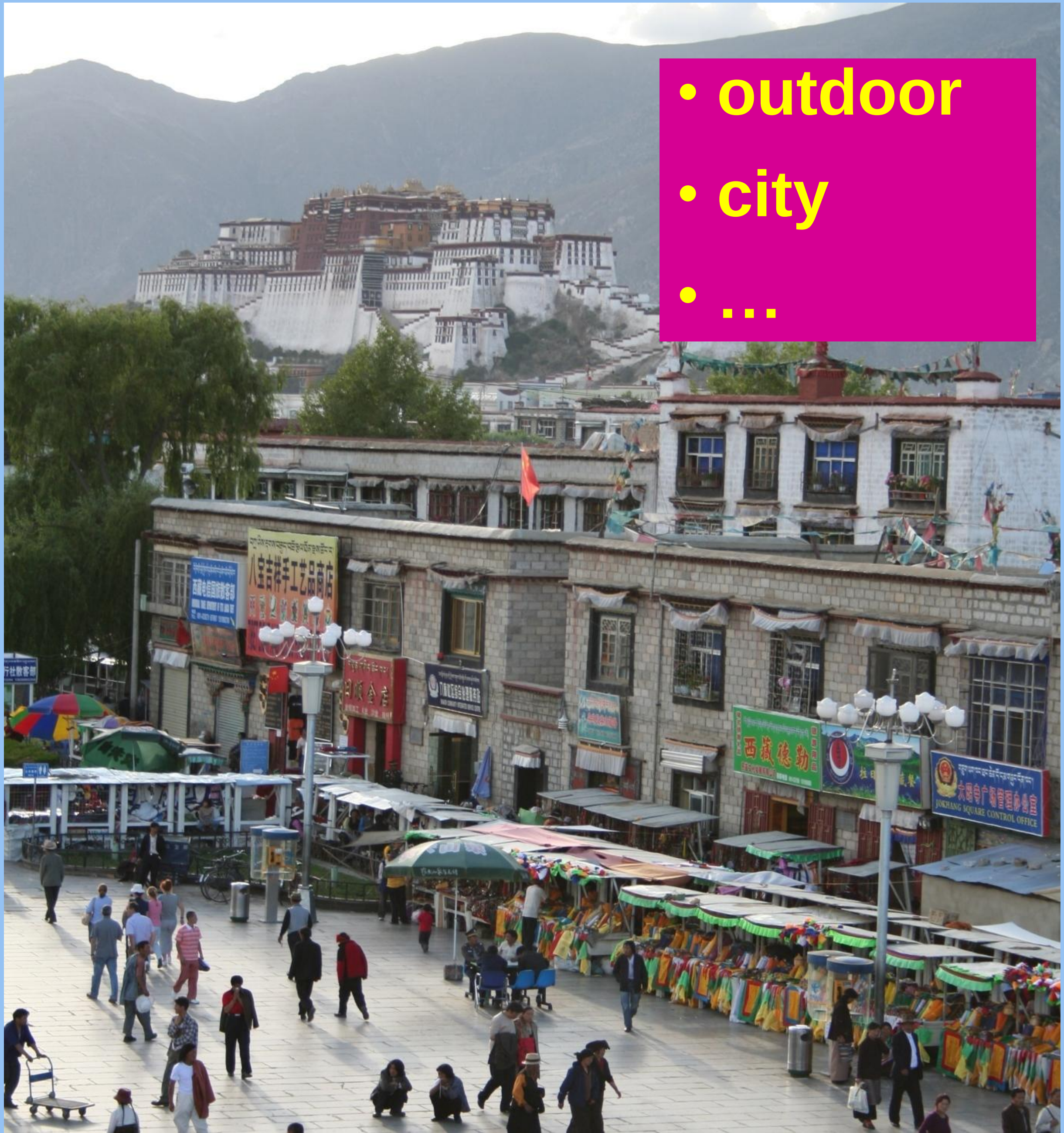
Detection: where are there people (if any)?

Identification: is that Potala Palace?

Object categorization



Scene and context categorization



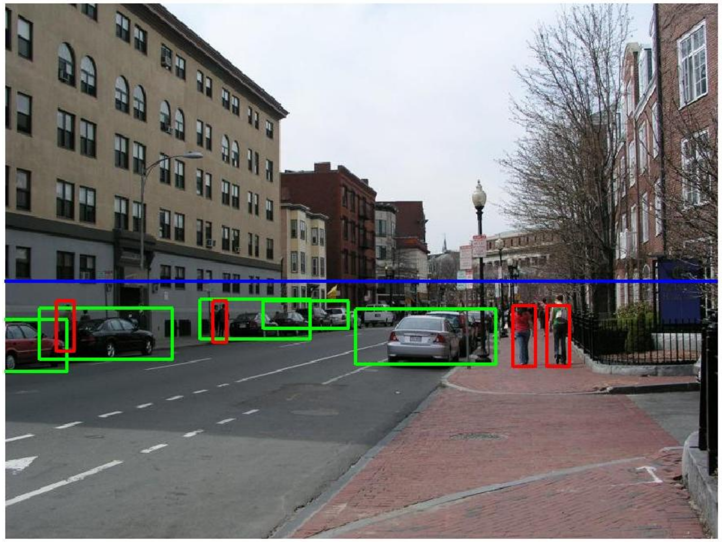
- **outdoor**
- **city**
- **…**

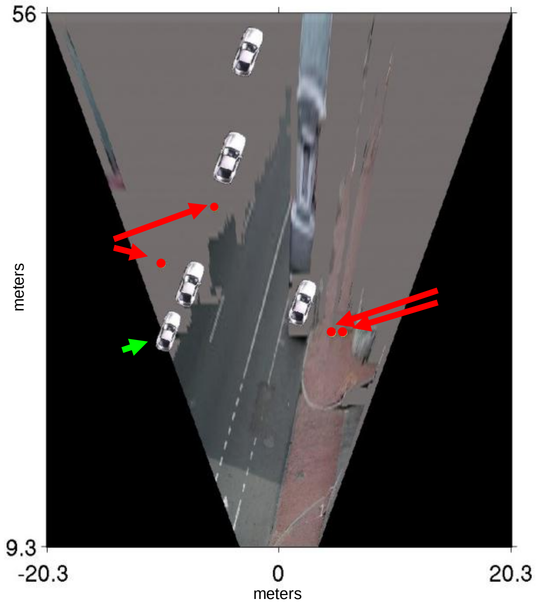
Applications: Photography

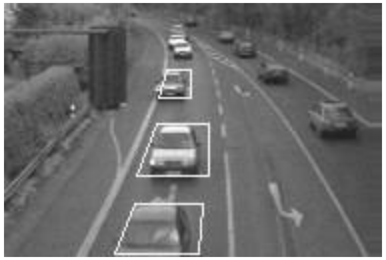
[Face priority AE] When a bright part of the face is too bright

Application: Assisted driving

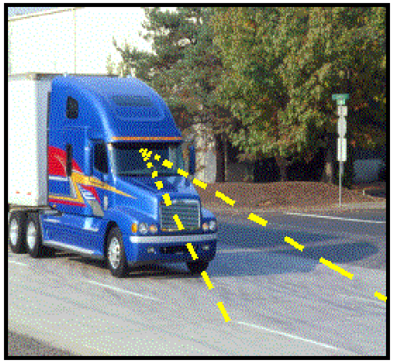
Pedestrian and car detection







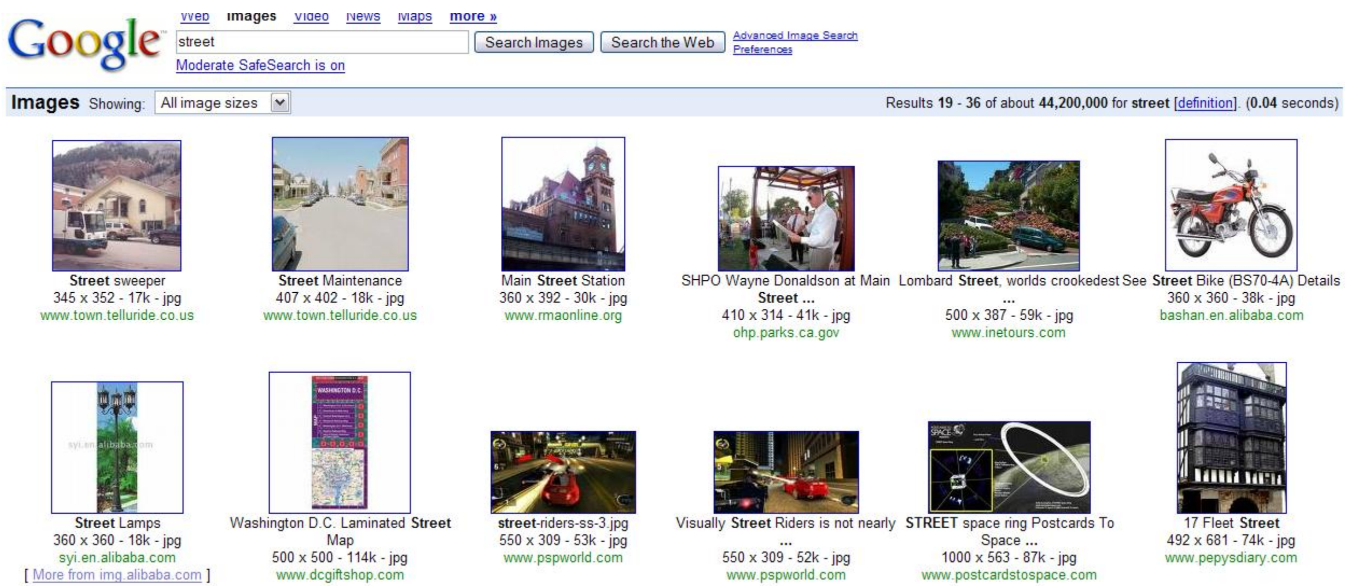
Lane detection



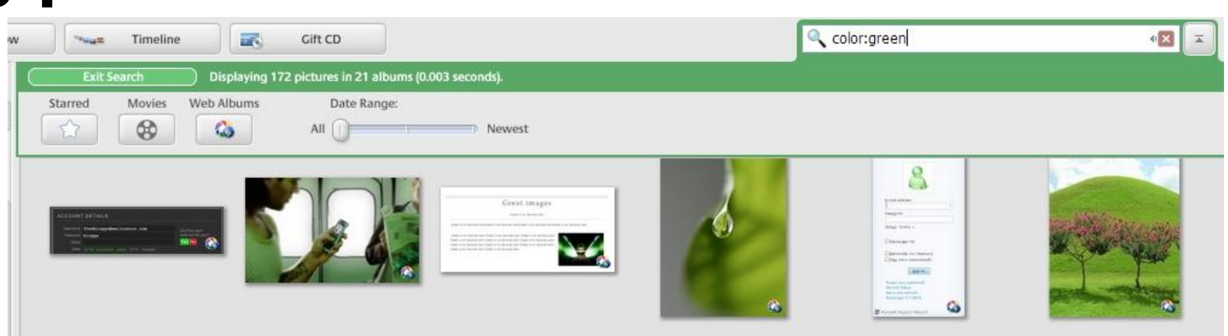
- Collision warning systems with adaptive cruise control,
- Lane departure warning systems,
- Rear object detection systems,

Application: Improving online search

Query:
STREET

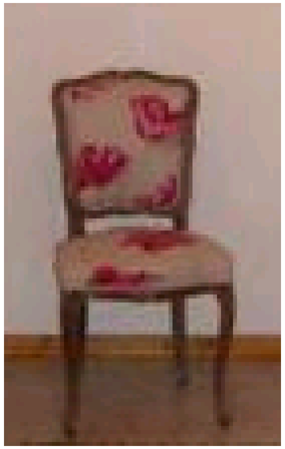


Organizing photo collections



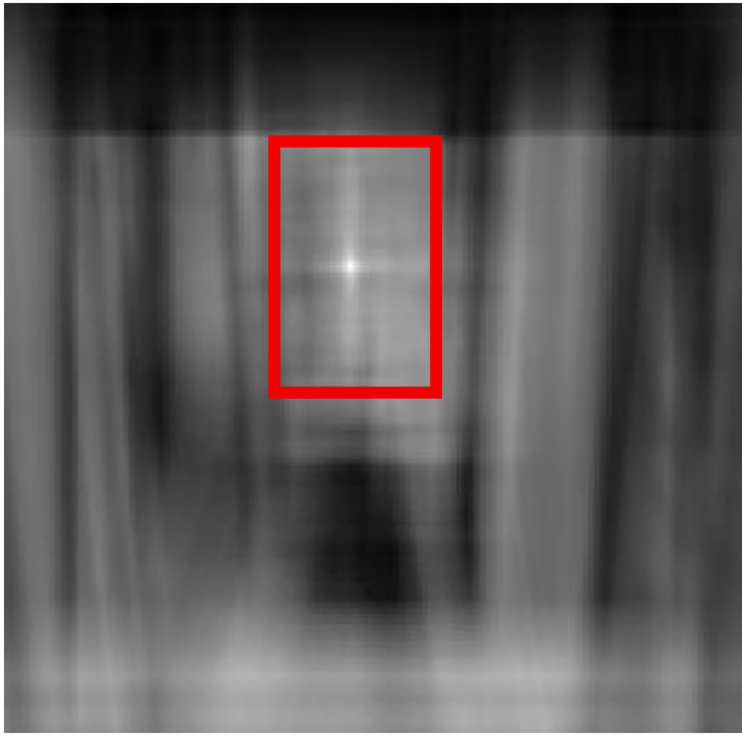
Object recognition
Is it really so hard?

This is a chair

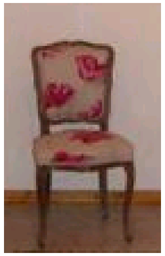


Find the chair in this image

Output of normalized correlation

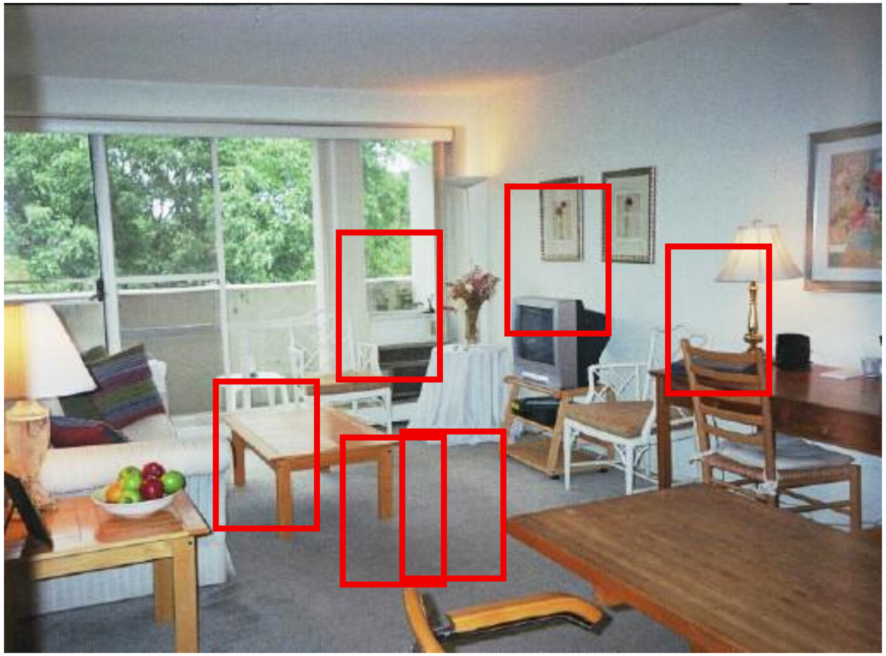


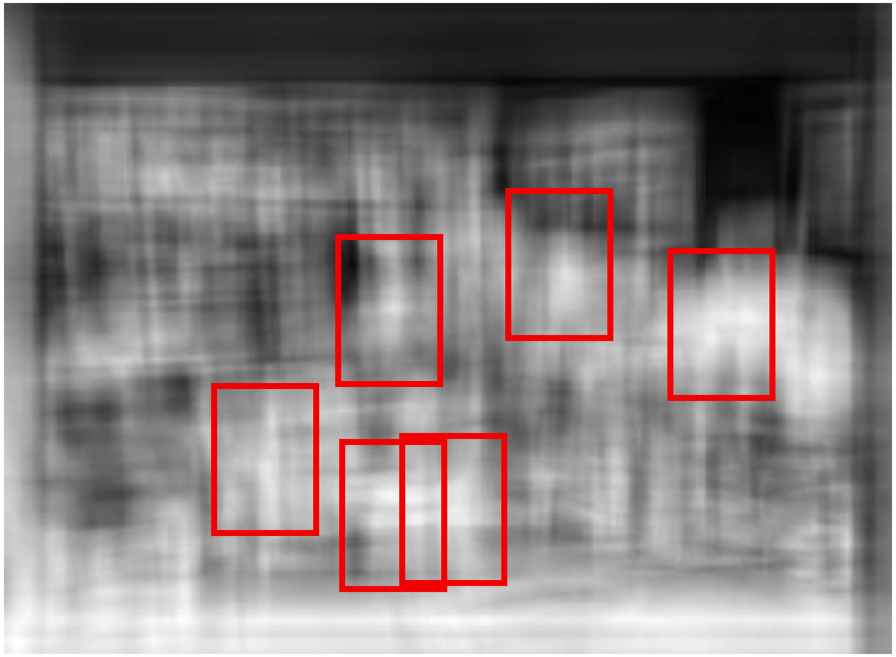
Slide: A. Torralba



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

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.

Slide: A. Torralba

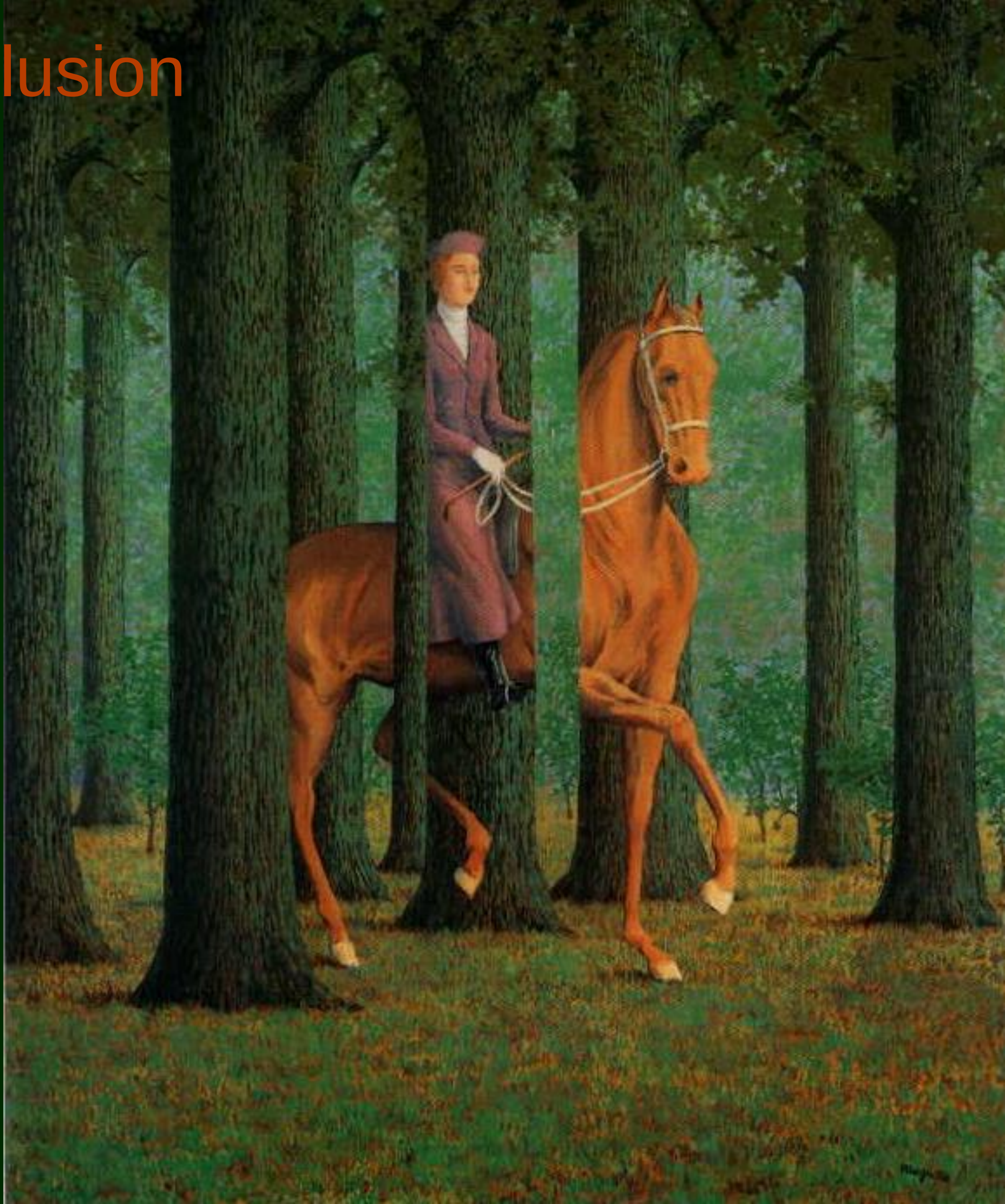
Challenges 1: view point variation

Michelangelo 1475-1564

Challenges 2: illumination

slide credit: S. Ullman

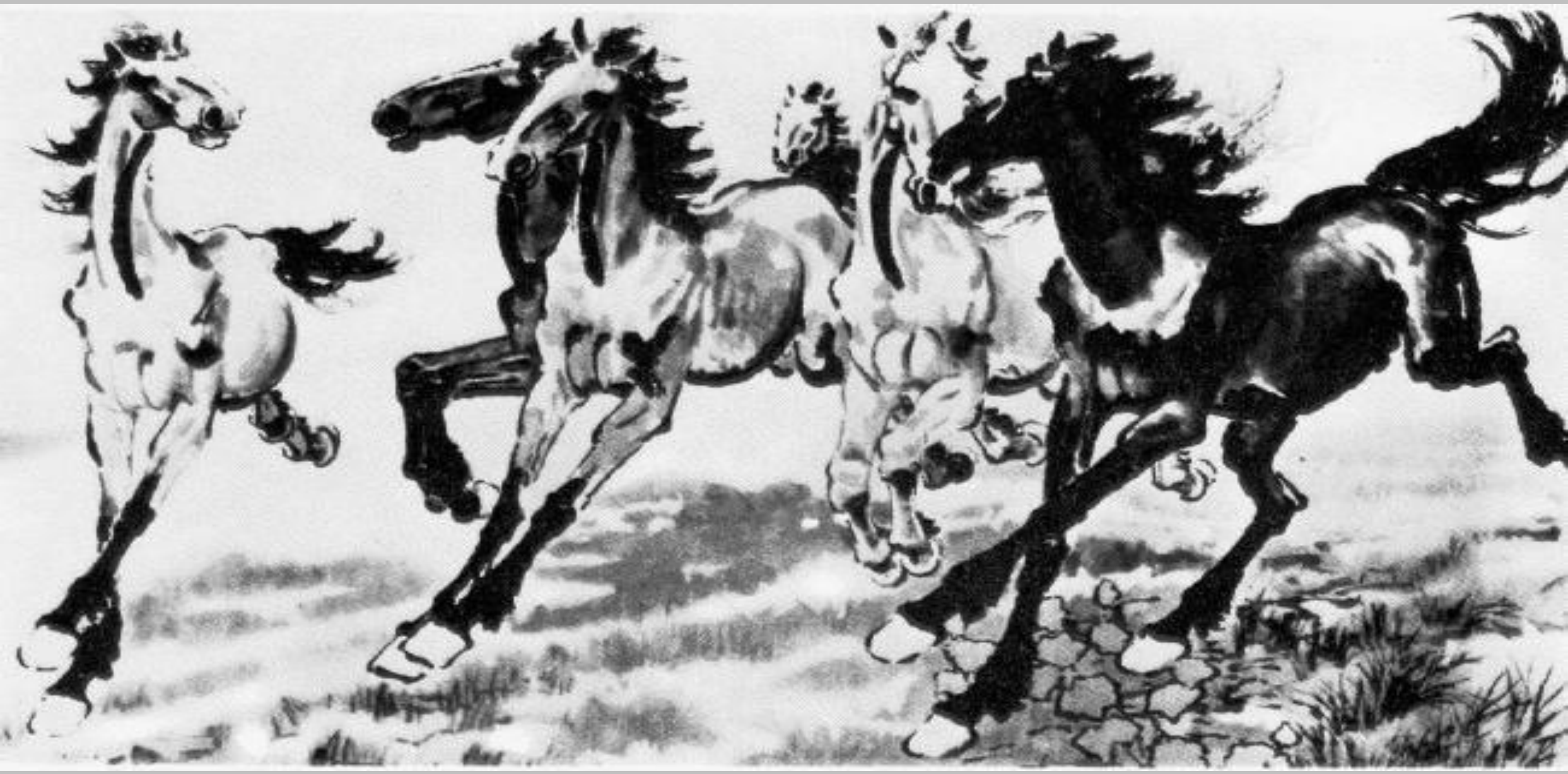
Challenges 3: occlusion



Magritte, 1957

Challenges 4: scale

Challenges 5: deformation

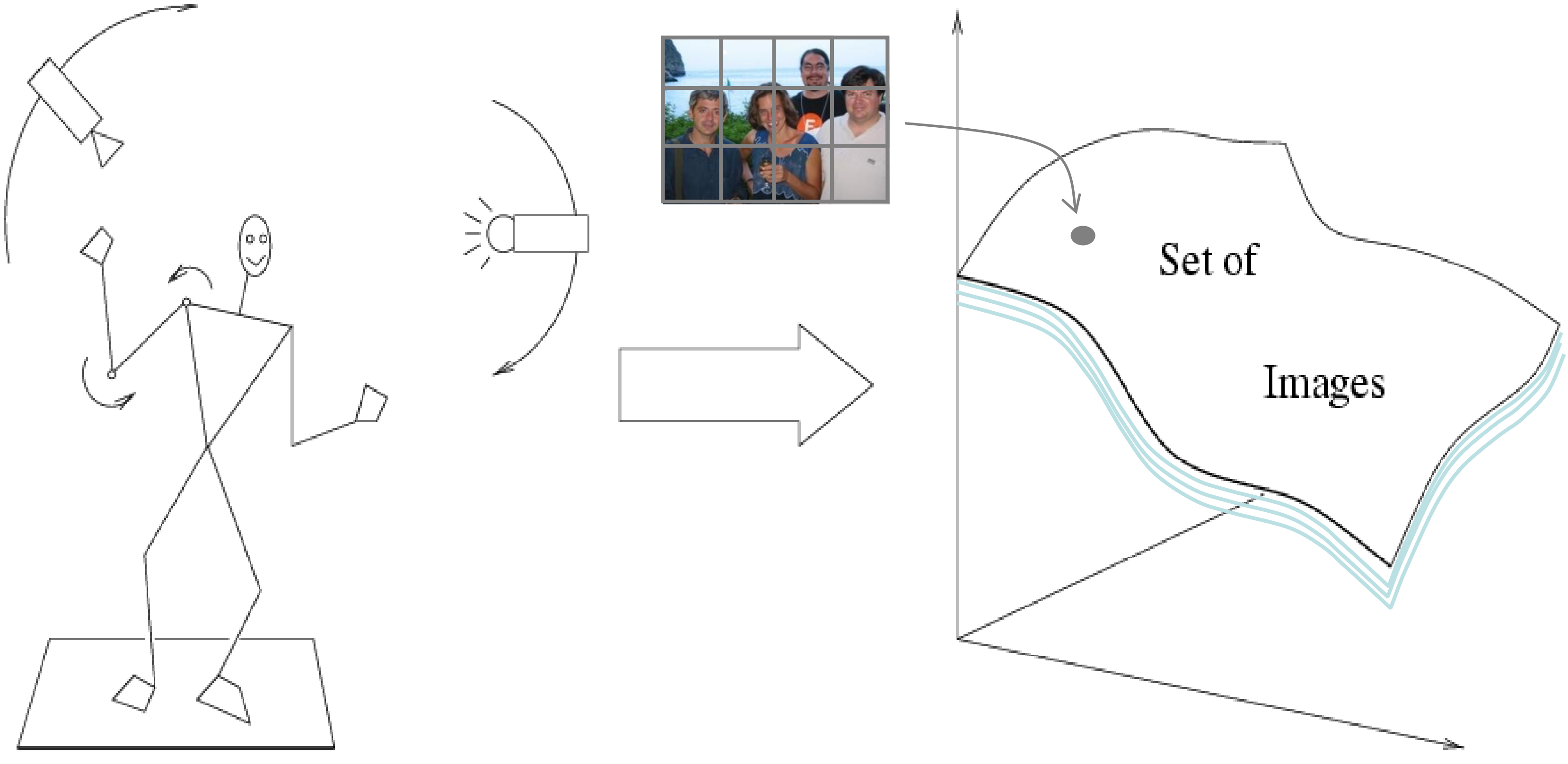


Xu, Beihong 1943

Challenges 6: background clutter

Klimt, 1913

Modeling variability



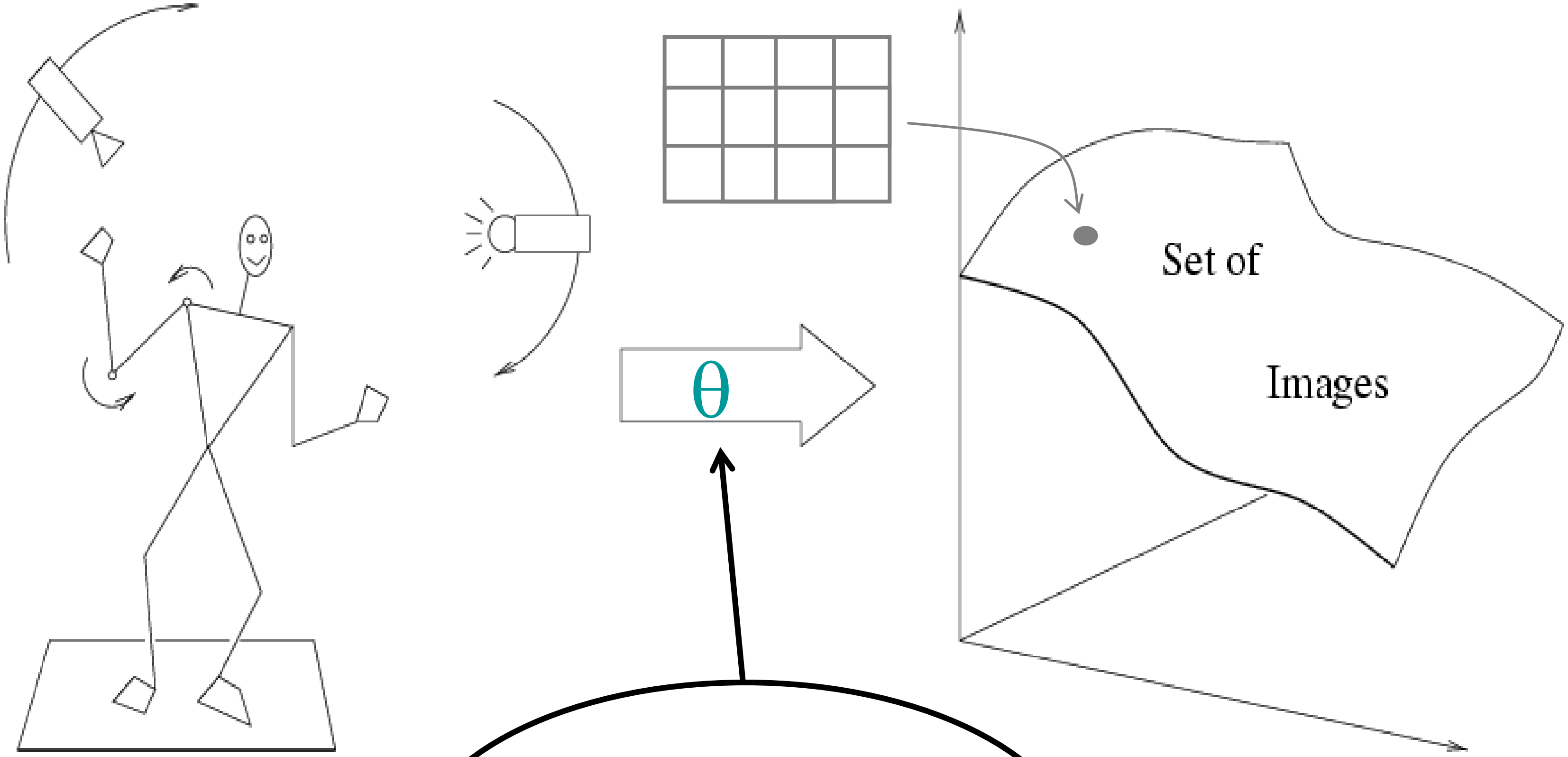
Variability: Camera position
Illumination
Internal parameters

 Within-class variations

Within-class variations

Timeline of recognition

- 1965-late 1980s: alignment, geometric primitives



Variability:

Camera position
Illumination
Internal parameters

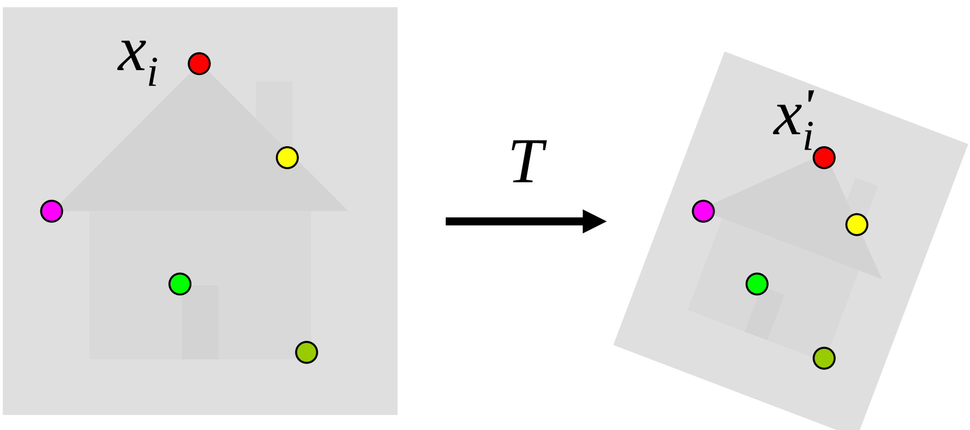
Alignment

Shape: assumed known

Roberts (1965); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

Recall: Alignment

- Alignment: fitting a model to a transformation between pairs of features (*matches*) in two images



Find transformation T
that minimizes

$$\sum_i \text{residual}(T(x_i), x_i')$$

Recognition as an alignment problem: Block world

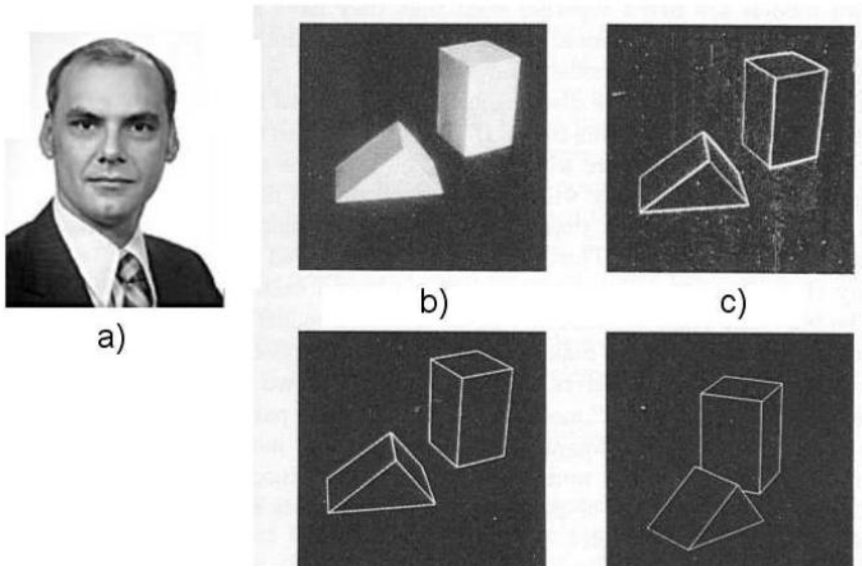


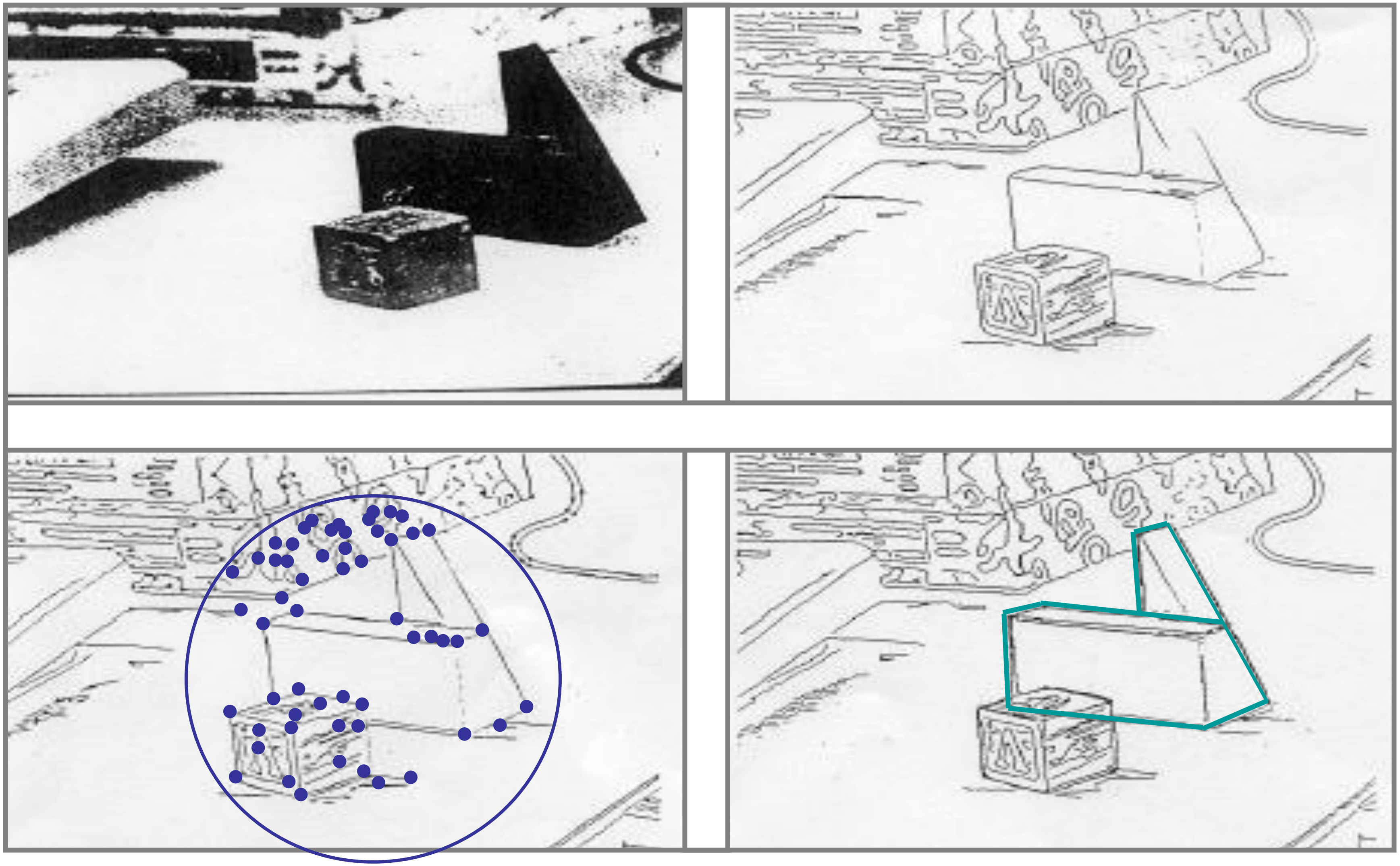
Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original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

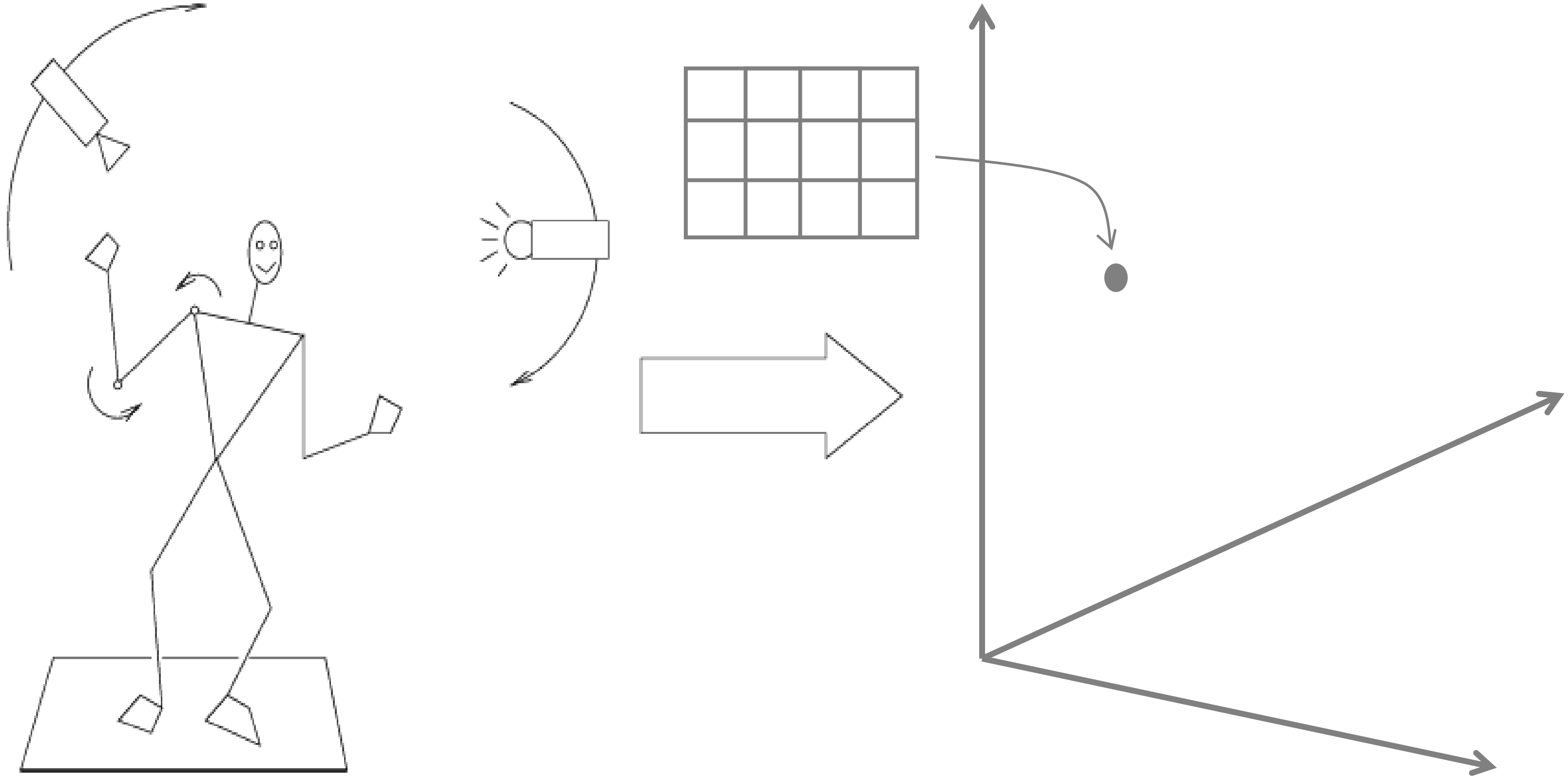
L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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

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

Alignment: Huttenlocher & Ullman (1987)



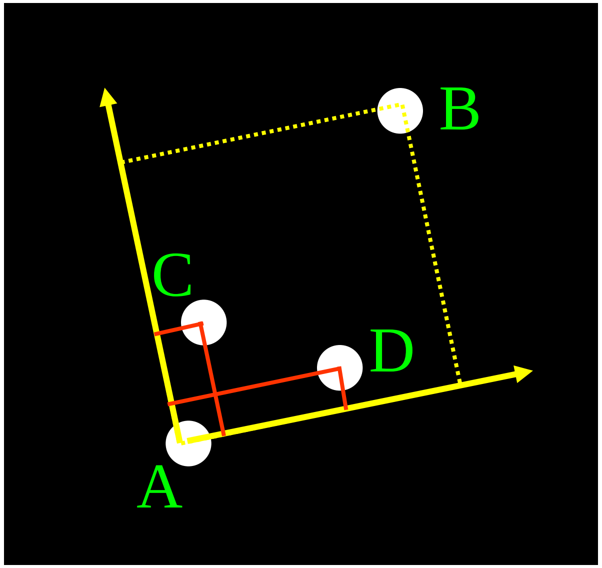


~~Variability~~ **Invariance to:**

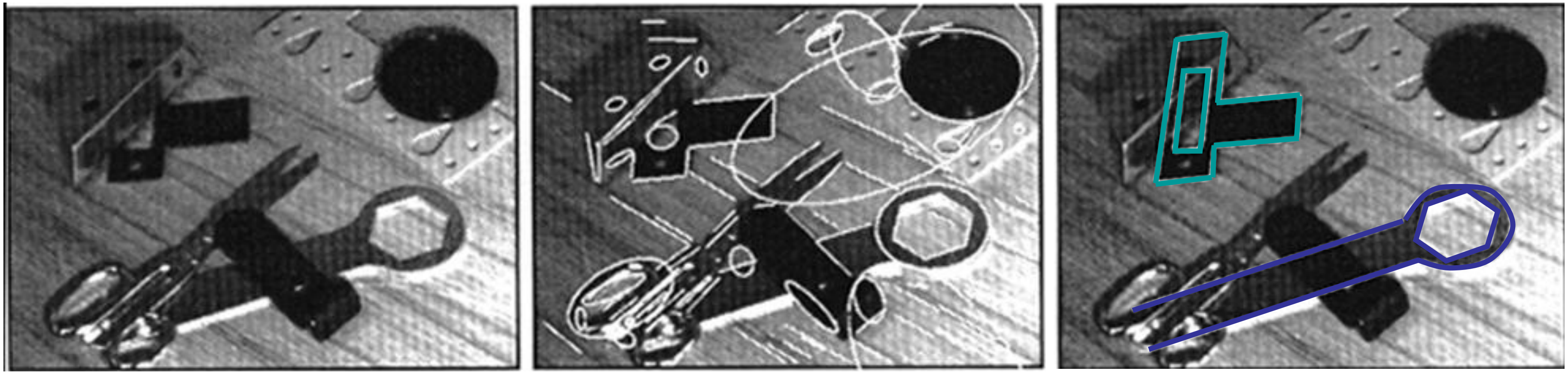
Camera position
Illumination
Internal parameters

Duda & Hart (1972); Weiss (1987); Mundy et al. (1992-94); Rothwell et al. (1992); Burns et al. (1993)

Example: invariant to similarity transformations computed from four points

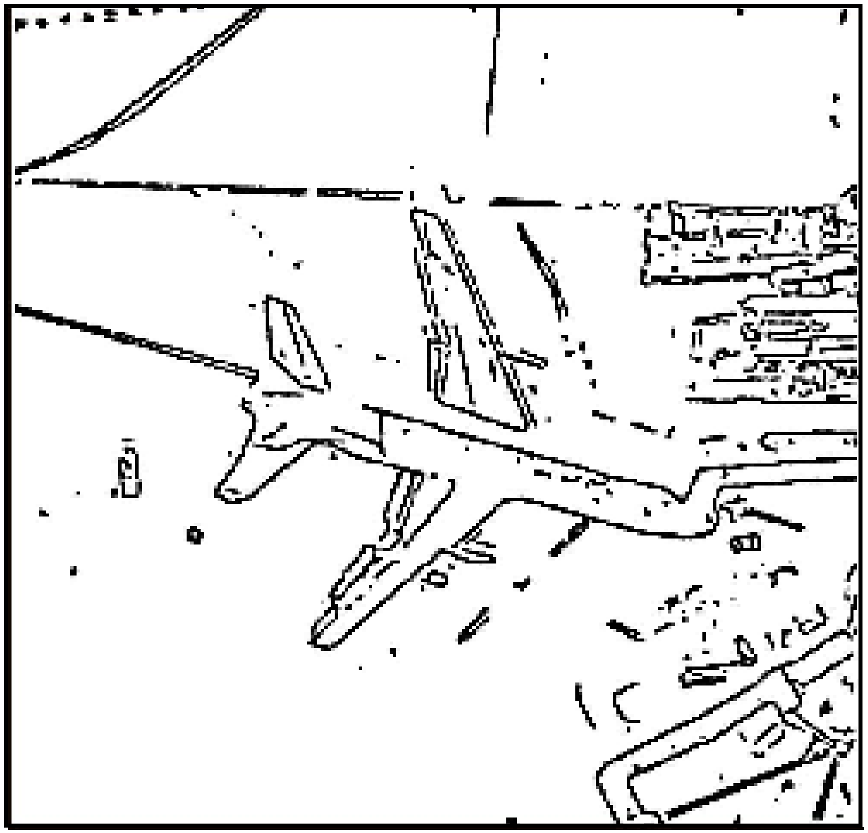


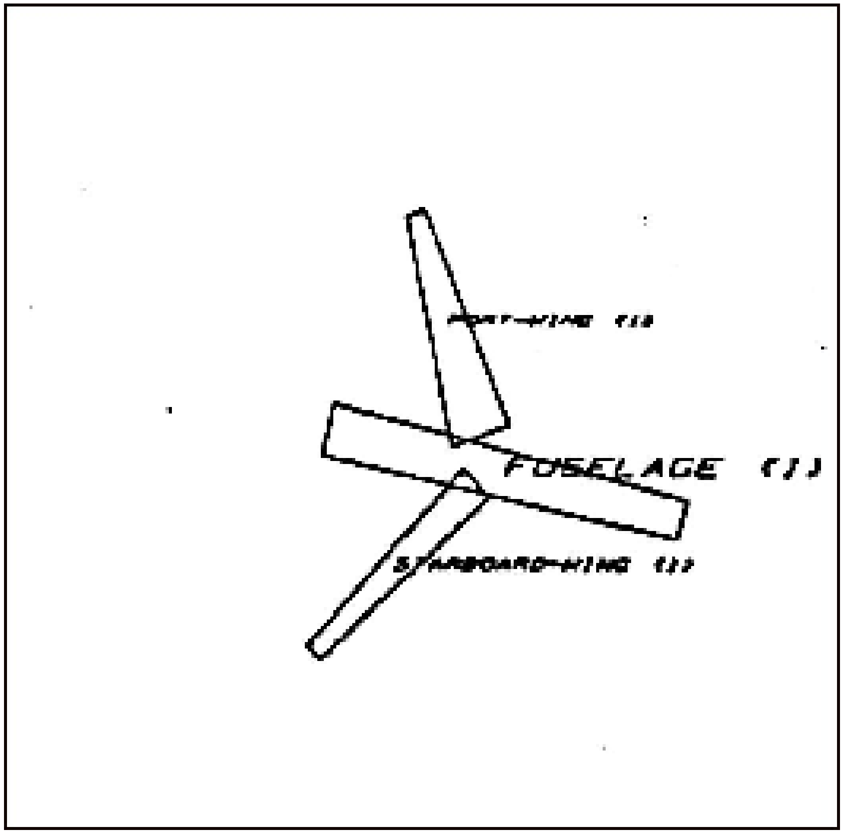
Projective invariants (Rothwell et al., 1992):



General 3D objects do not admit monocular viewpoint invariants (Burns et al., 1993)

Representing and recognizing object categories is harder...





ACRONYM (Brooks and Binford, 1981)

Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

Binford and generalized cylinders

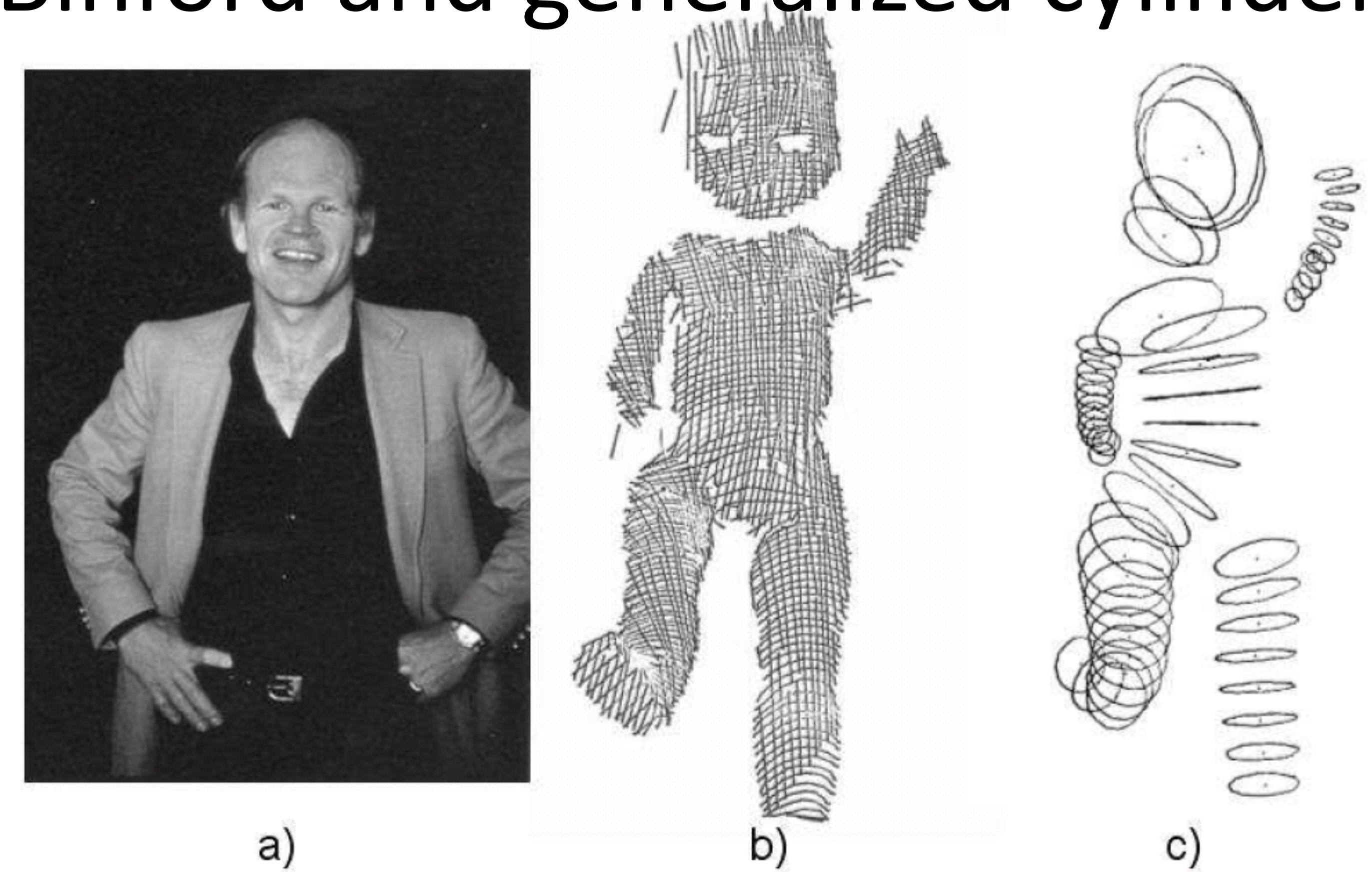
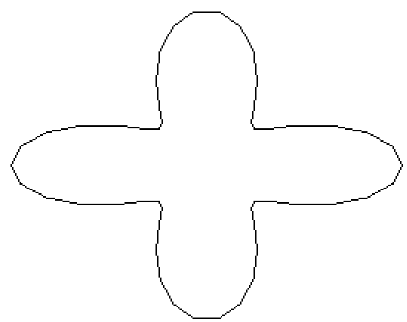


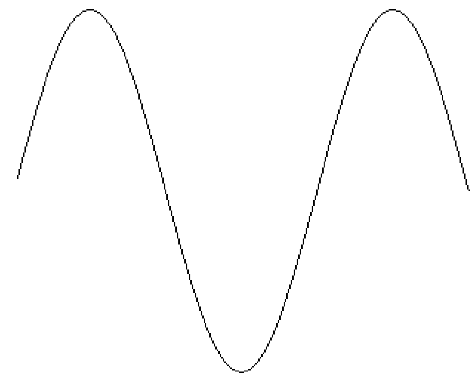
Fig. 3. The representation of objects by assemblies of generalized cylinders. a) Thomas Binford. b) A range image of a doll. c) The resulting set of generalized cylinders. (b) and c) are taken from Agin [1] with permission.)

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

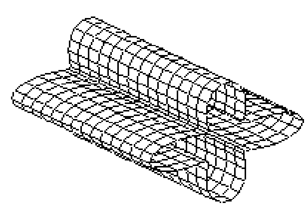
Binford and generalized cylinders



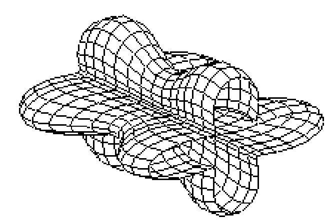
(a) Cross section.



(b) Sweeping rule.

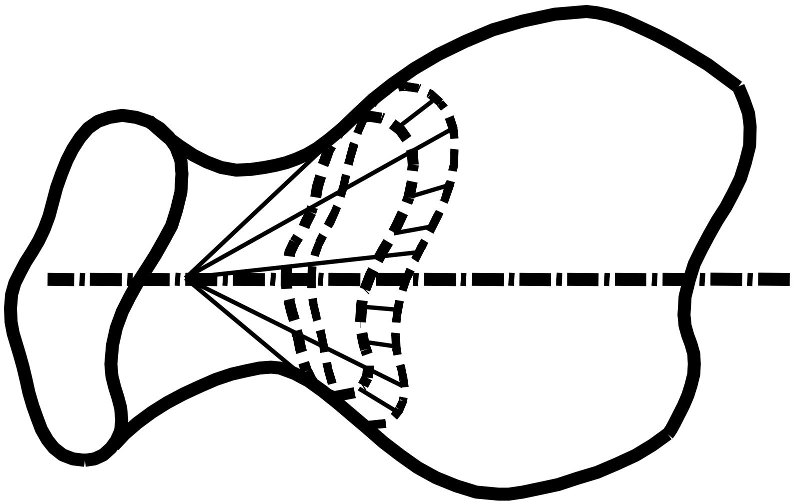


(c) True cylinder

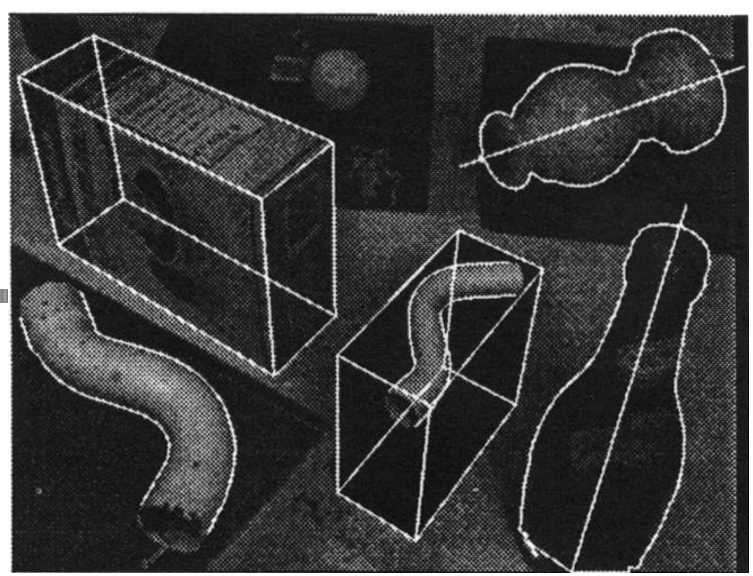


(d) Generalized cylinder

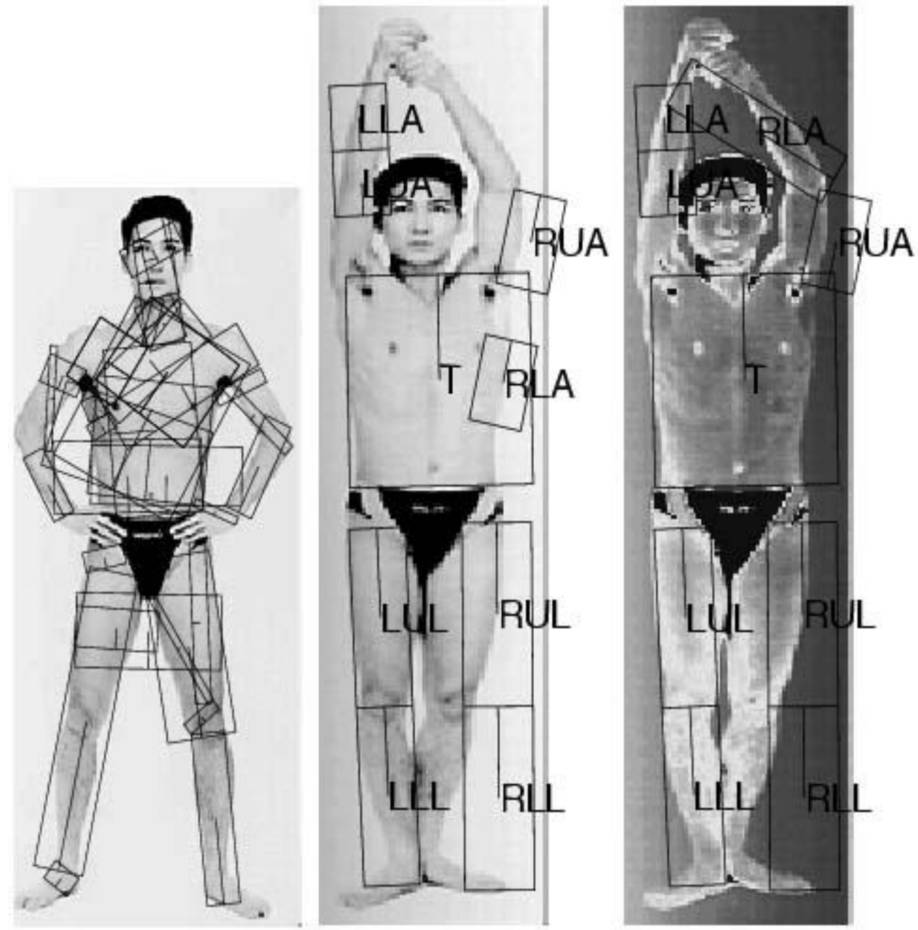
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



Zisserman et al. (1995)



Forsyth (2000)

Recognition by components

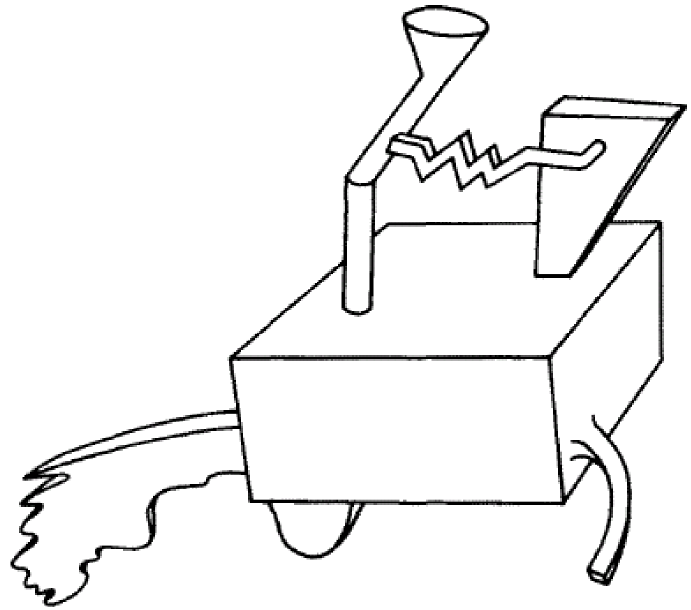
Irving Biederman
Recognition-by-Components: A Theory of Human Image Understanding.
Psychological Review, 1987.

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.

Stages of processing

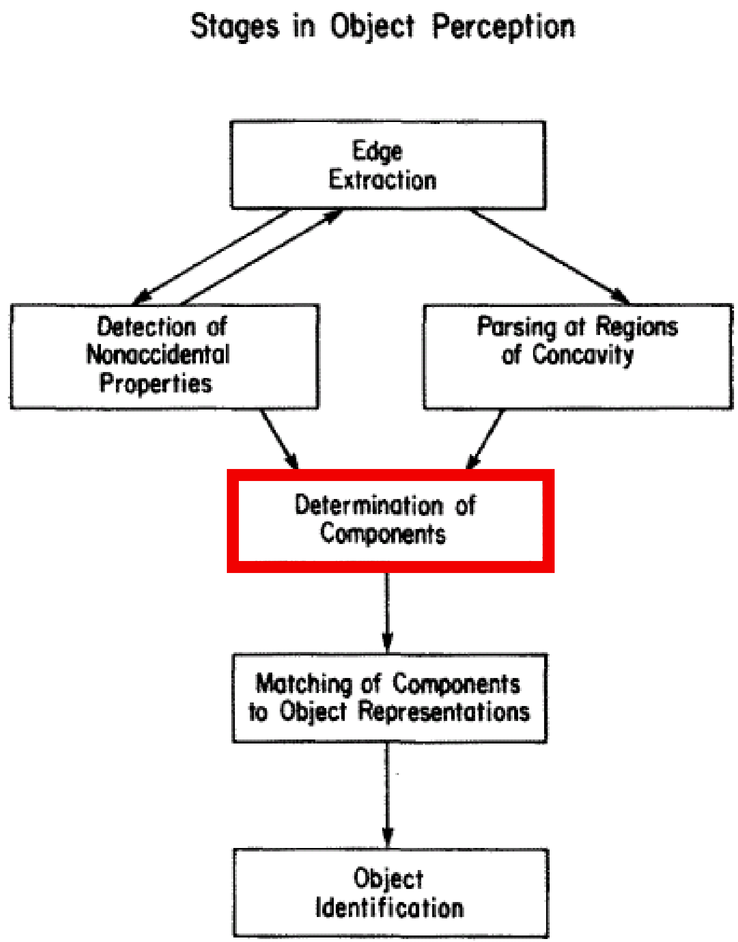


Figure 2. Presumed processing stages in object recognition.

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

Examples:

- Colinearity
- Smoothness
- Symmetry
- Parallelism
- Cotermination

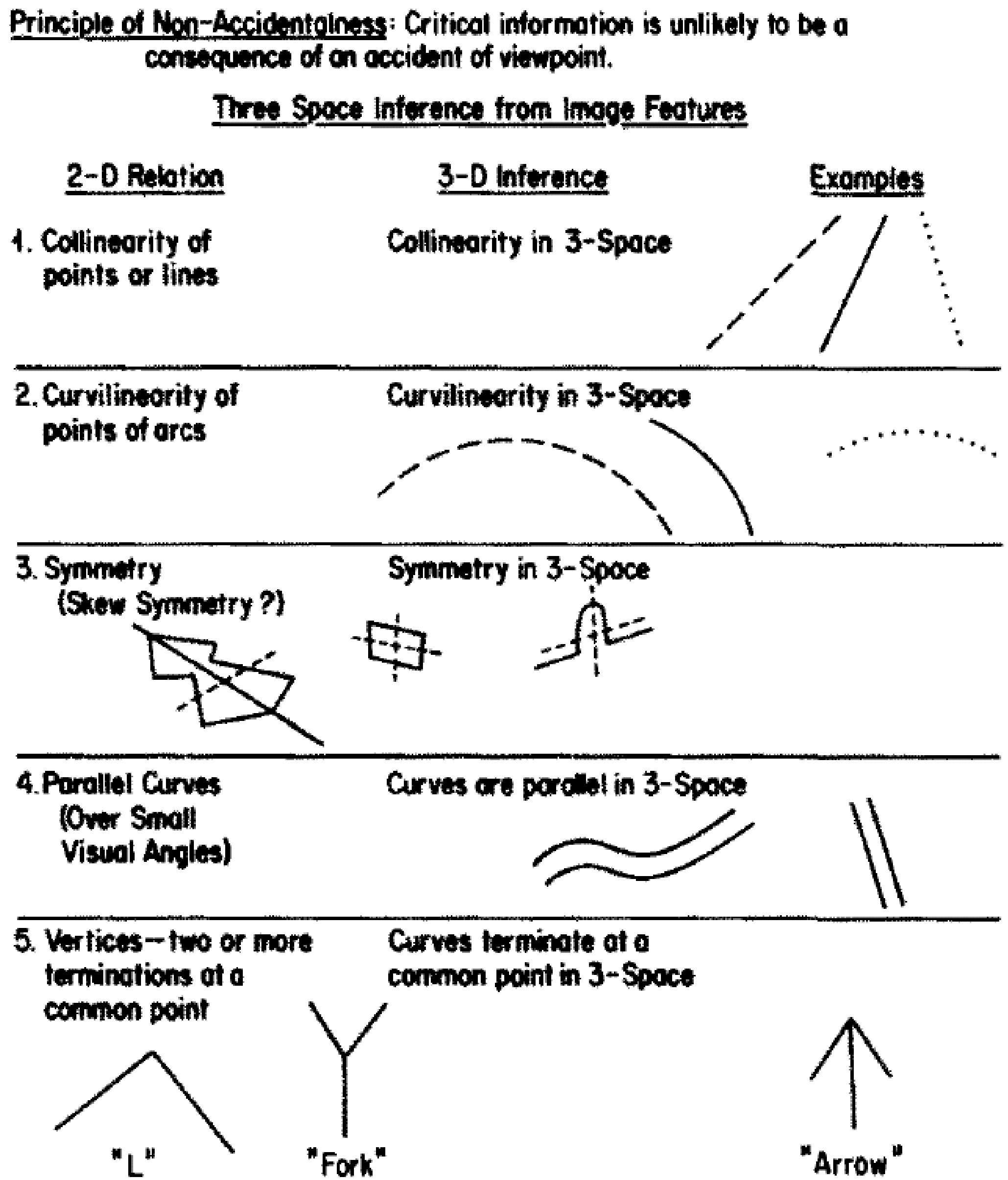
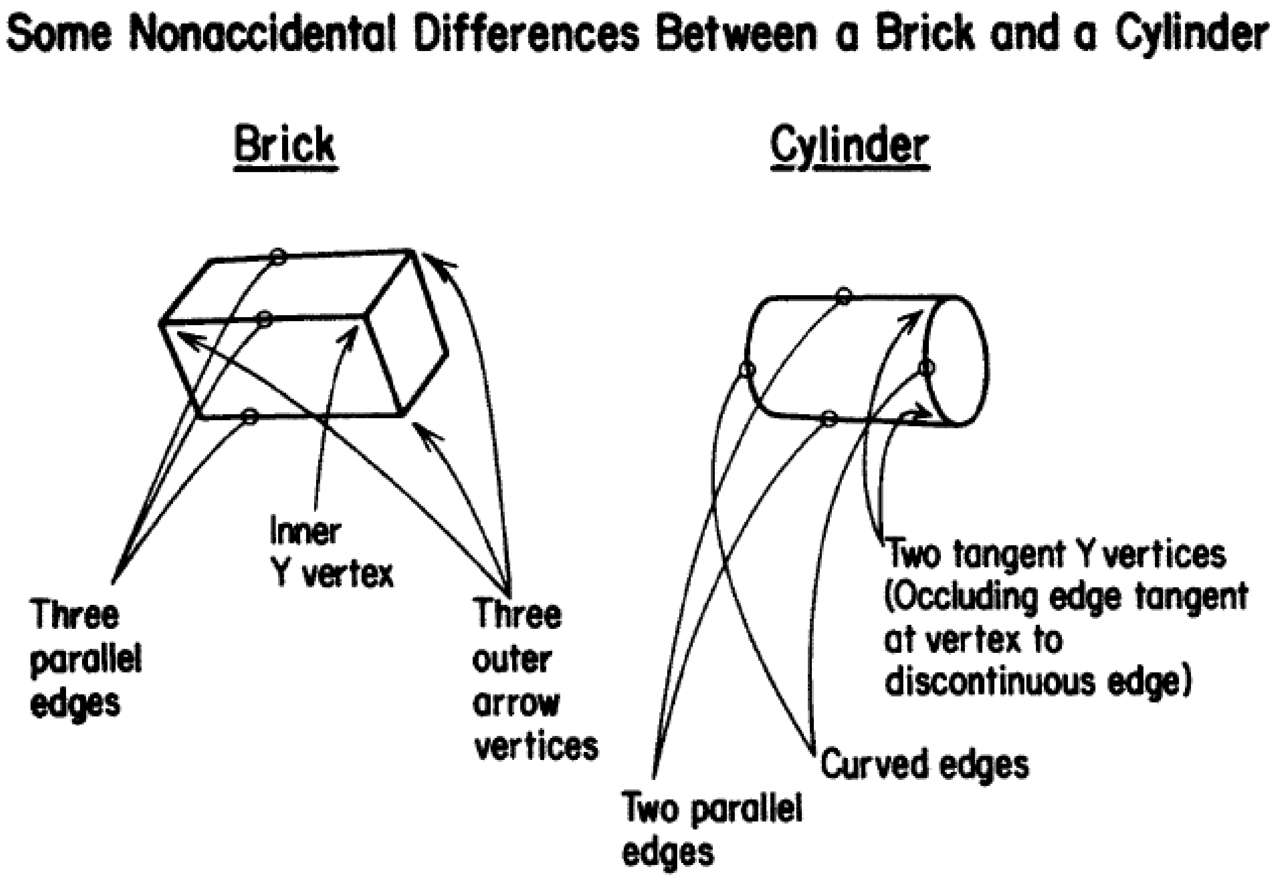
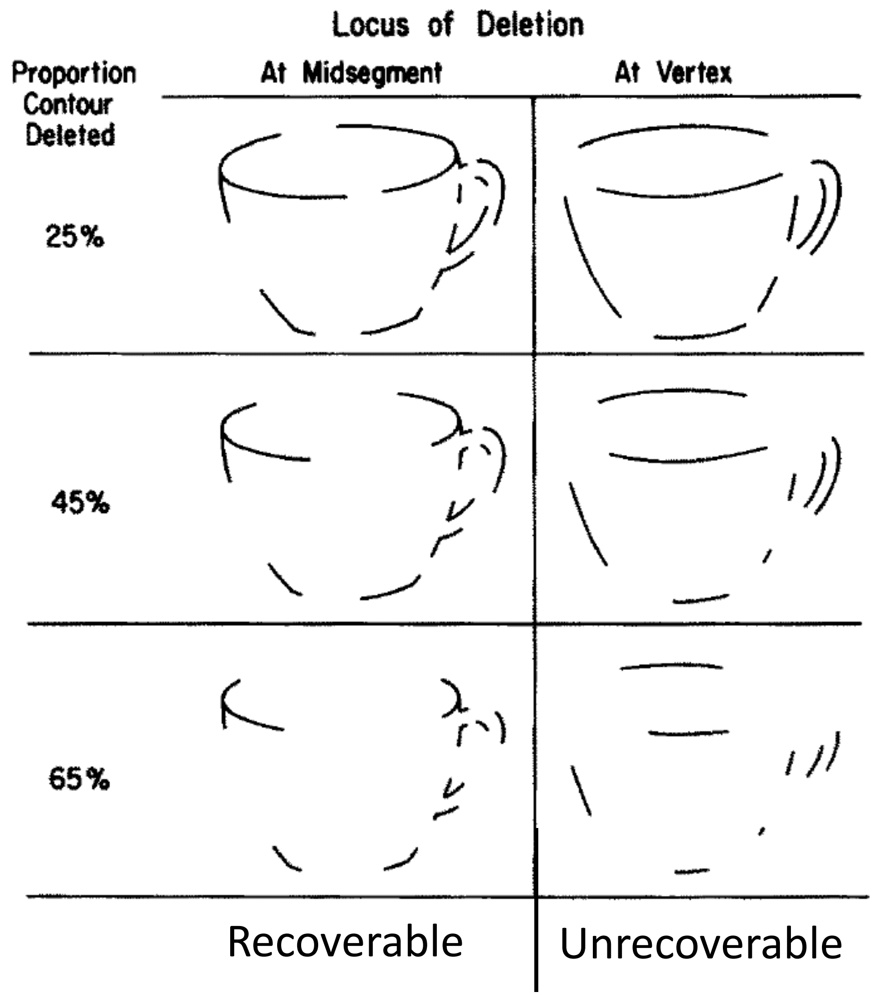


Figure 4. Five nonaccidental relations. (From Figure 5.2, *Perceptual organization and visual recognition* [p. 77] by David Lowe. Unpublished doctorial dissertation, Stanford University. Adapted by permission.)



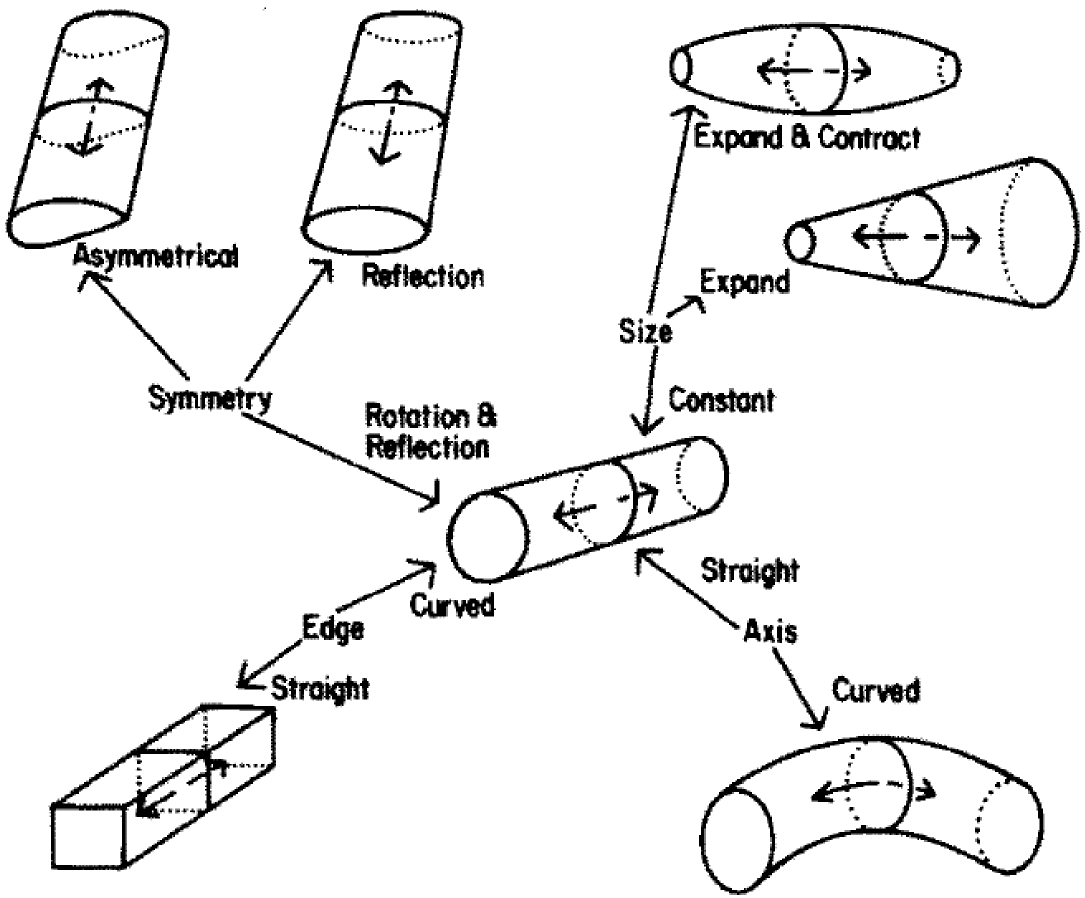
The high speed and accuracy of determining a given nonaccidental relation {e.g., whether some pattern is symmetrical) should be contrasted with performance in making absolute quantitative judgments of variations in a single physical attribute,
such as length of a segment or degree of tilt or curvature.

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."

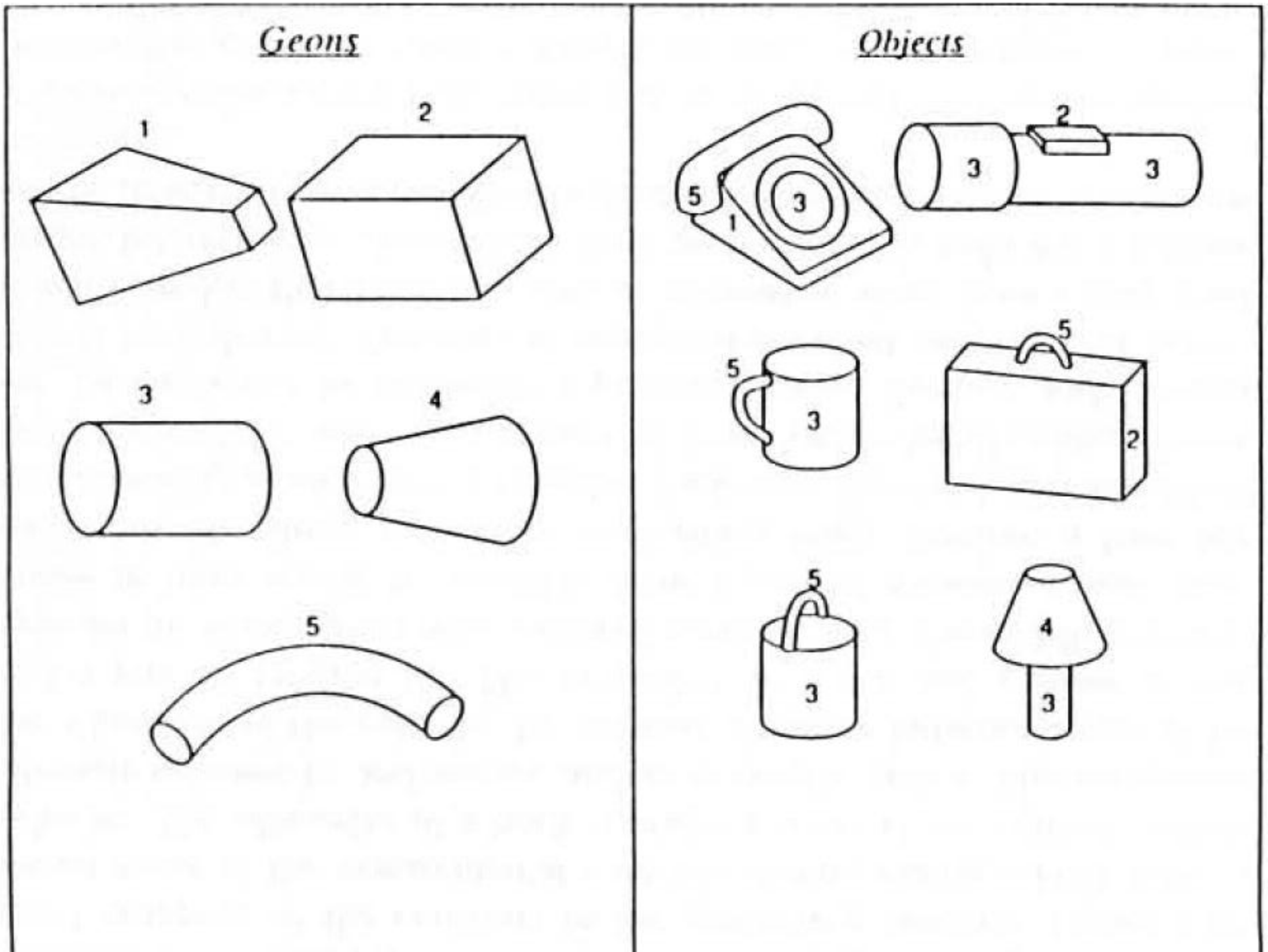
From generalized cylinders to GEONS



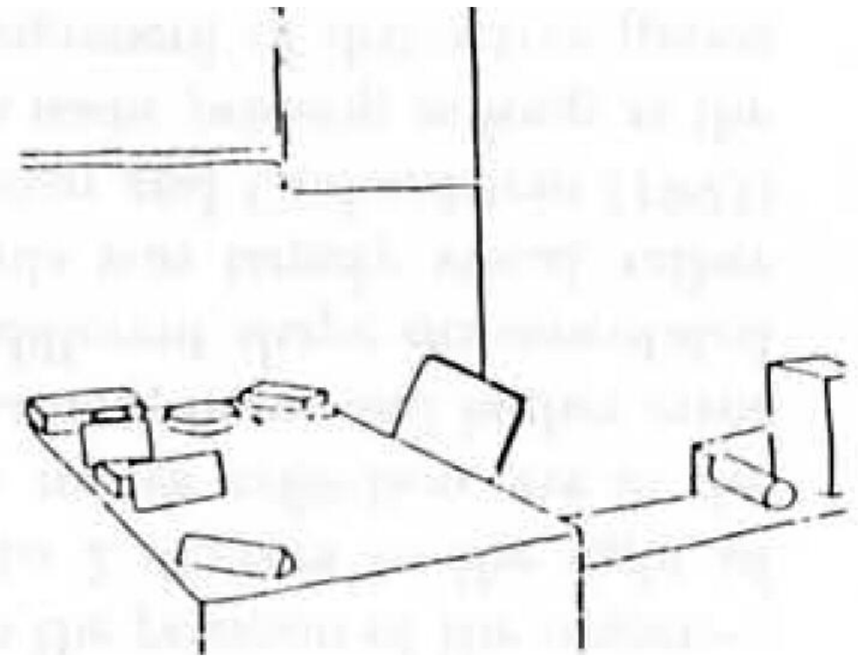
"From variation over only two or three levels in the nonaccidental relations of four attributes of generalized cylinders, a set of 36 GEONS can be generated."

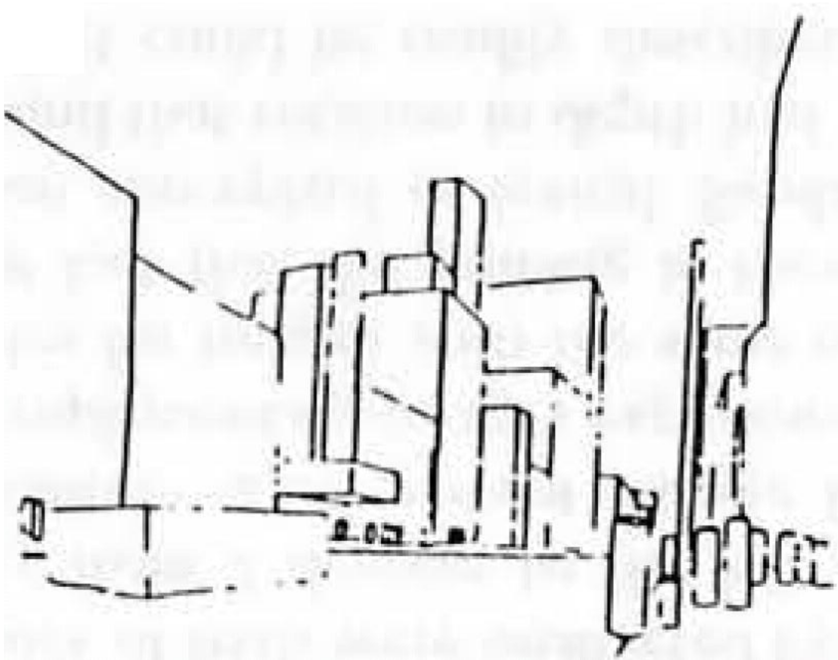
Geons represent a restricted form of generalized cylinders.

Objects and their geons



Scenes and geons





Mezzanotte & Biederman

The importance of spatial arrangement

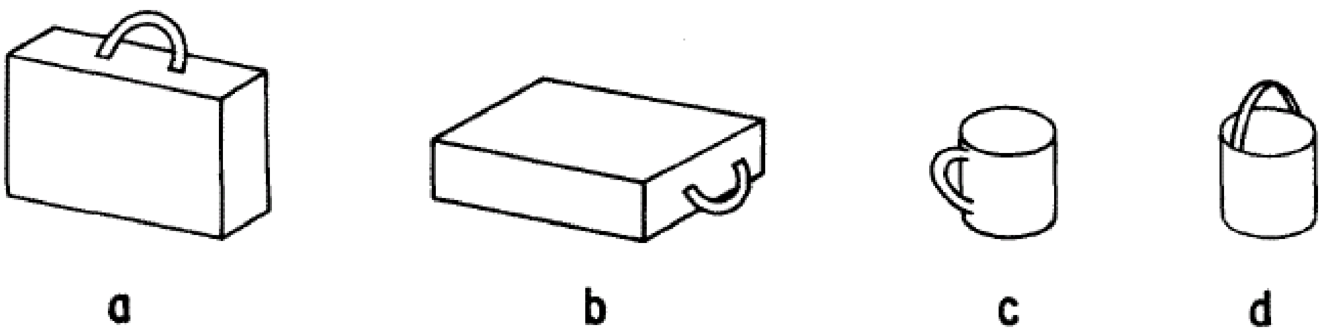
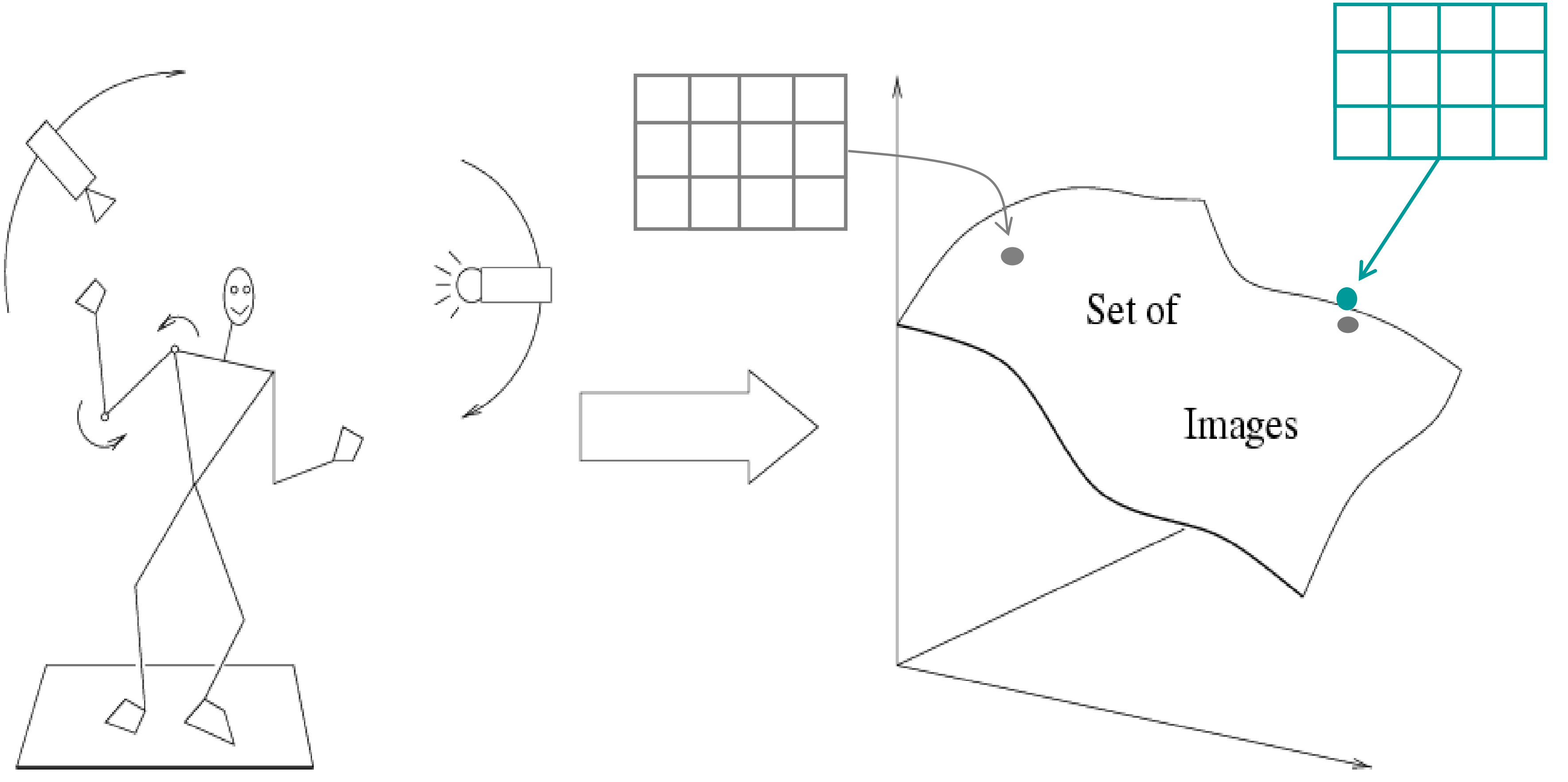


Figure 3. Different arrangements of the same components can produce different objects.

Timeline of recognition

- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods

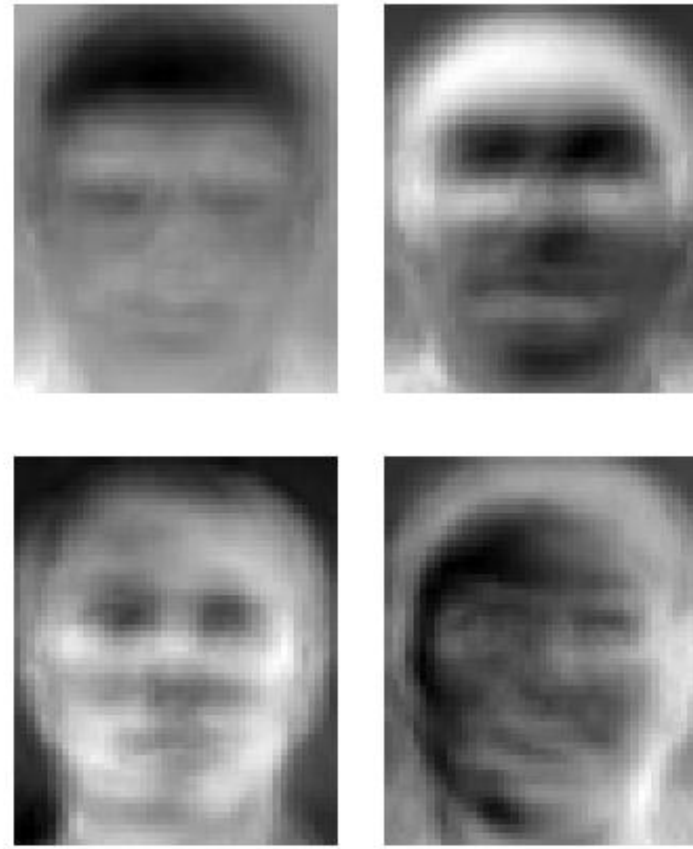


Empirical models of image variability

Appearance-based techniques

Turk & Pentland (1991); Murase & Nayar (1995); etc.

Eigenfaces (Turk & Pentland, 1991)

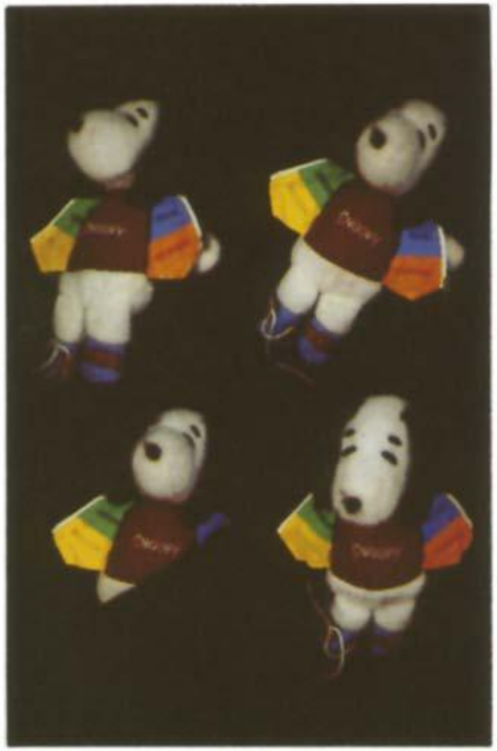


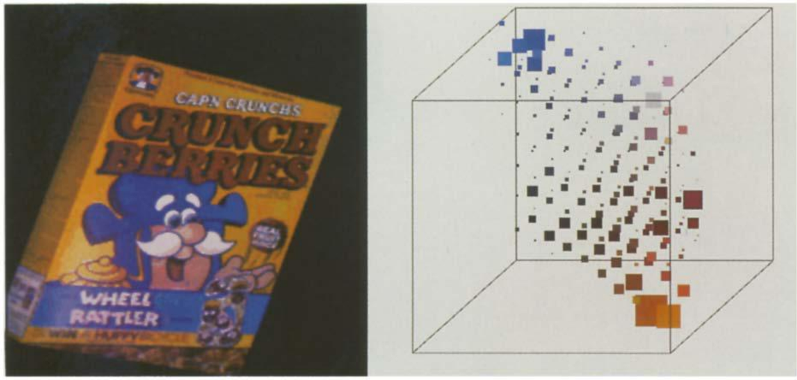
Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

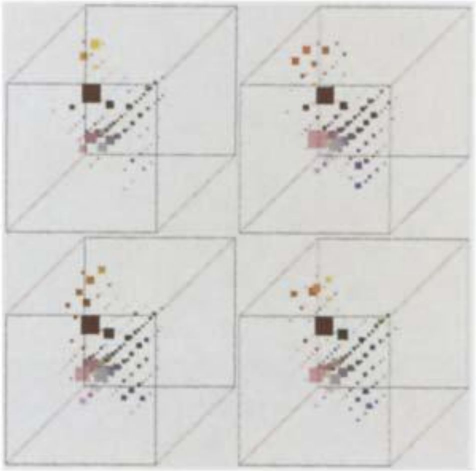
Eigenfaces

Explain on whiteboard

Color Histograms

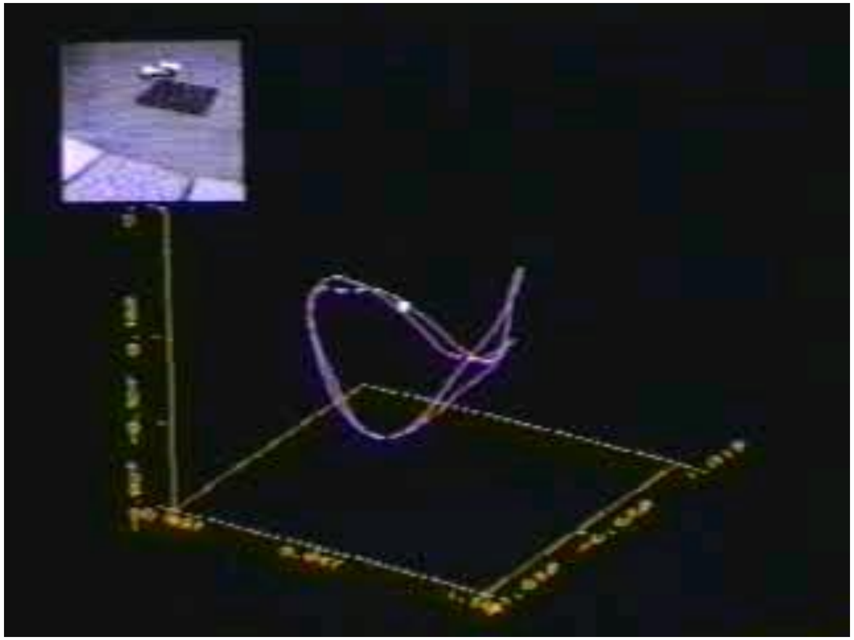






Swain and Ballard, Color Indexing, IJCV 1991.

Appearance manifolds



H. Murase and S. Nayar, Visual learning and recognition of 3-d objects from appearance, IJCV 1995

Limitations of global appearance models

- Can work on relatively simple patterns

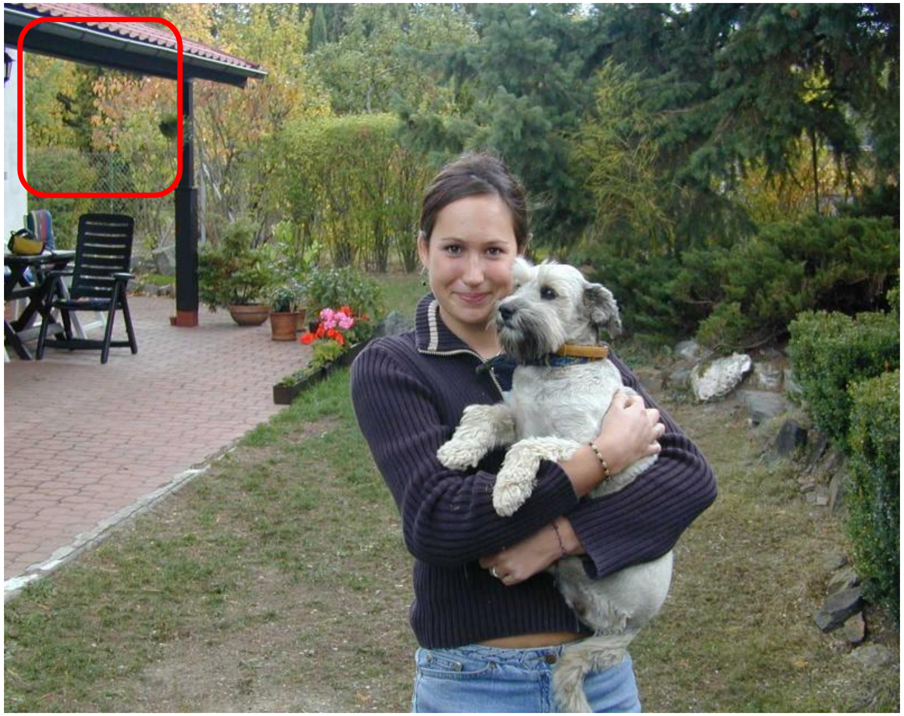
- Not robust to clutter, occlusion, lighting changes

Timeline of recognition

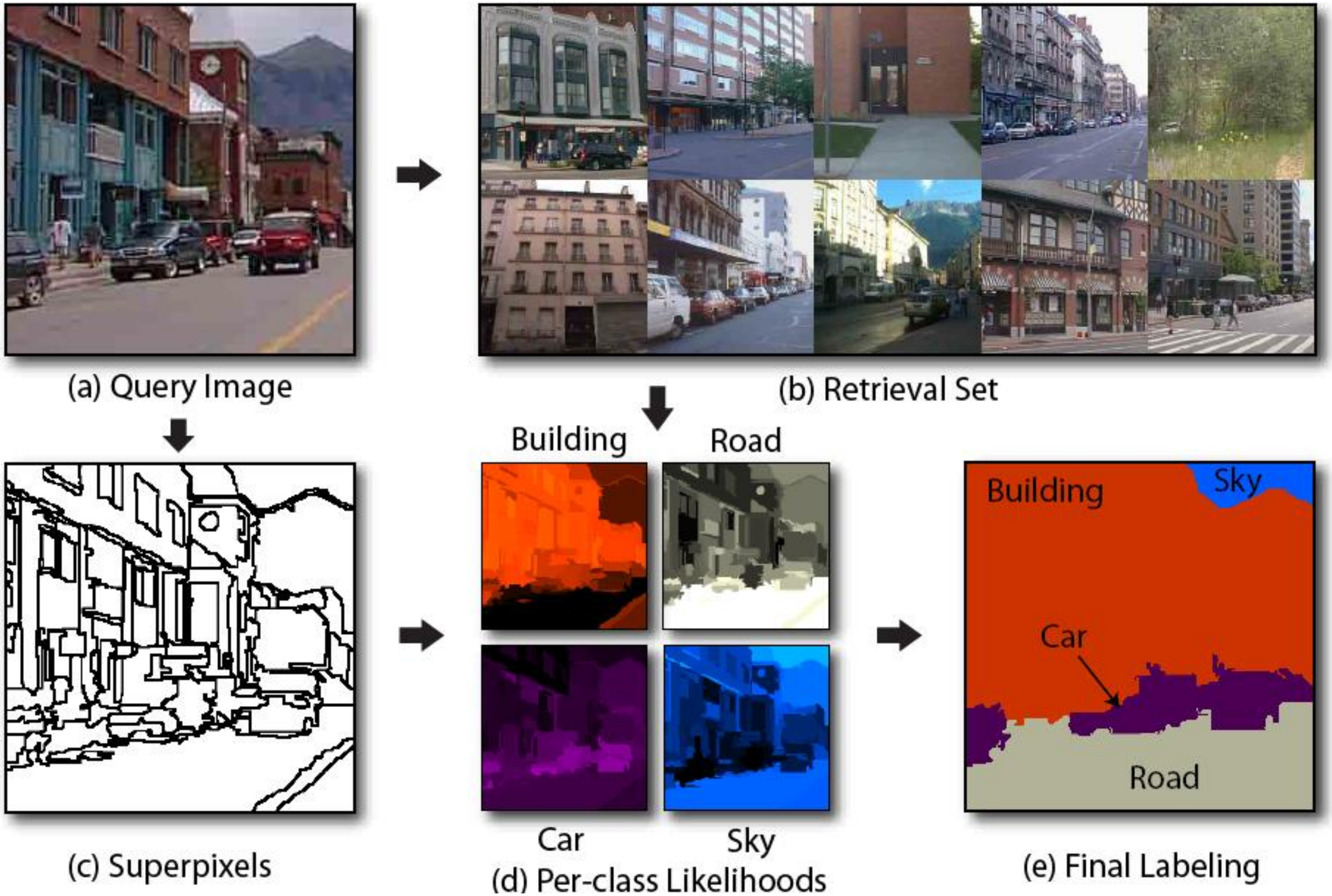
- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods
- Mid-late 1990s: sliding window approaches

Sliding window approaches

- Classify each window separately
- Scale / orientation range to search over

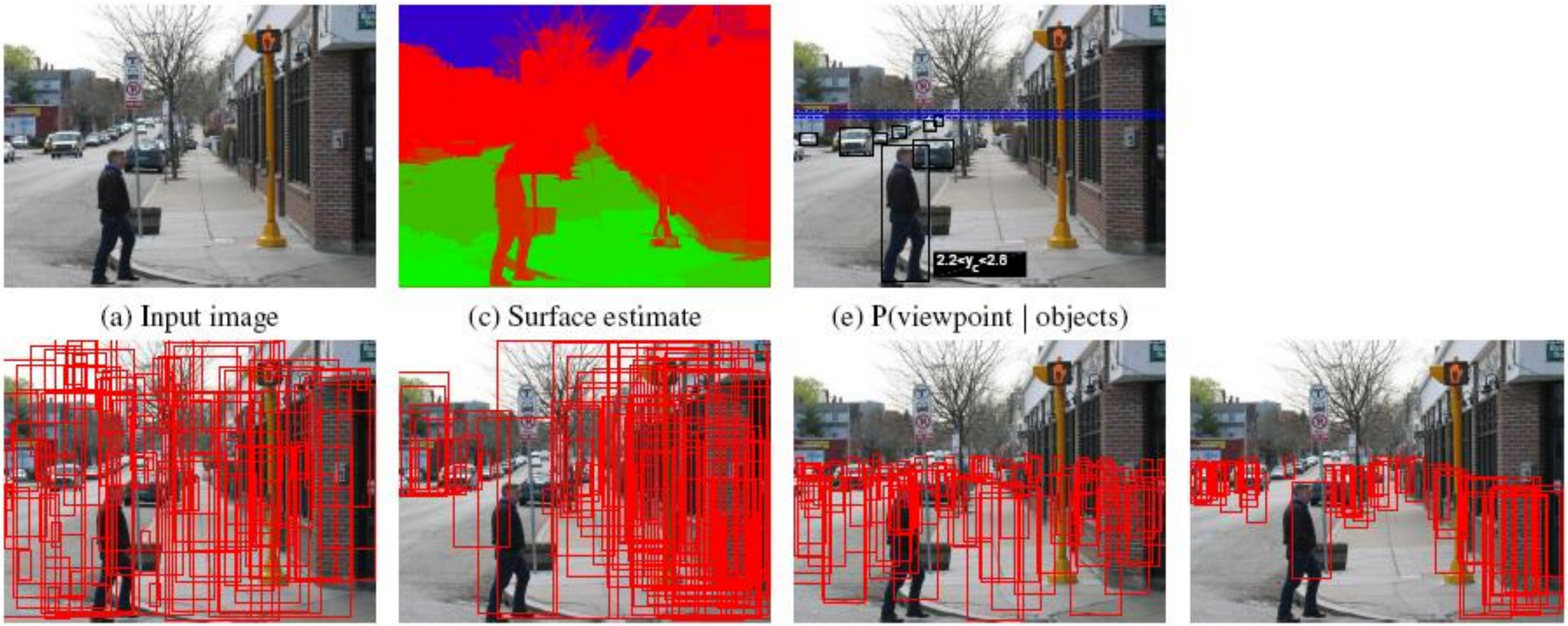


Scene-level context for image parsing



J. Tighe and S. Lazebnik, ECCV 2010 submission

Geometric context



(a) Input image (c) Surface estimate (e) P(viewpoint | objects)

(b) P(person) = uniform (d) P(person | geometry) (f) P(person | viewpoint) (g) P(person|viewpoint,geometry)

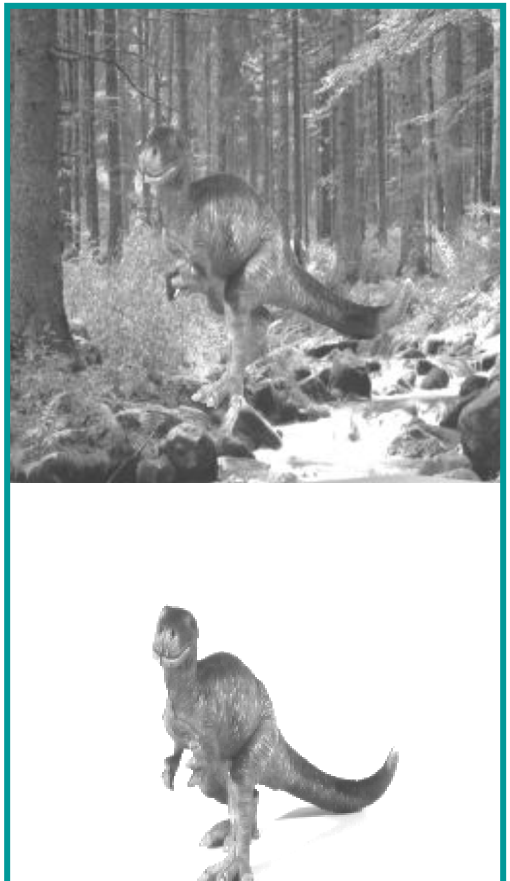
D. Hoiem, A. Efros, and M. Herbert. [Putting Objects in Perspective.] CVPR 2006.

Timeline of recognition

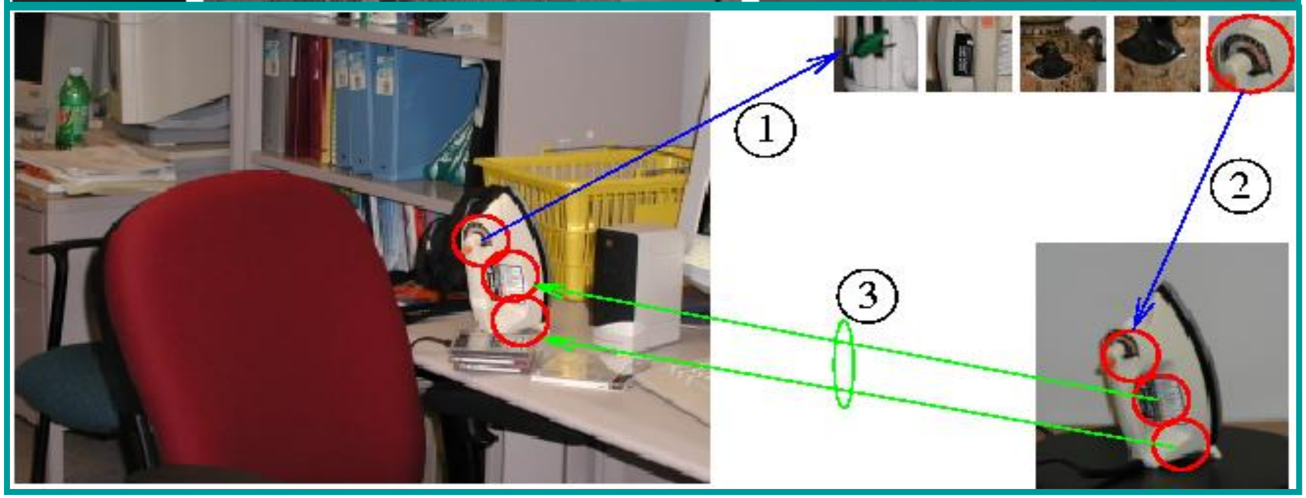
- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods
- Mid-late 1990s: sliding window approaches
- Late 1990s: feature-based methods

Local features

Combining *local* appearance, spatial constraints, invariants, and classification techniques from machine learning.

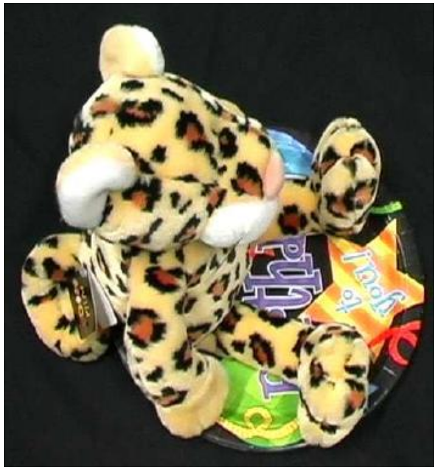


Schmid & Mohr'97

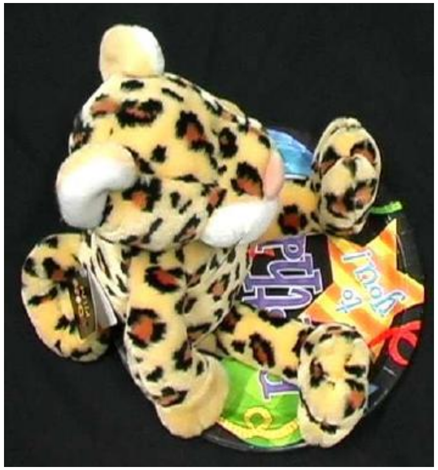


Mahamud & Hebert'03

Specific Object Recognition



Specific Object Recognition



Specific Object Recognition Application

Evolution Robotics Retail™ Our vision, your profit

Home | Products | Our Company | News & Events | CONTACT US >

Products

- LaneHawk BOB
  - Benefits
  - Store Integration
  - Demonstration Video
  - Stopping BOB Loss
- LaneHawk InCart
- LaneHawk Enterprise Manager
- ShelfHawk
- TunnelHawk
- Core Technologies

LaneHawk® BOB

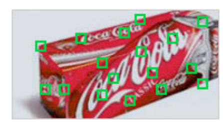
LaneHawk® BOB is a loss-prevention solution that turns bottom-of-basket (BOB) losses into profits in real time.

WATCH DEMO VIDEO >

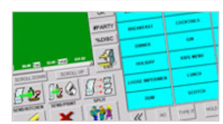
It starts with a visual scanner that recognizes items without having to read the UPC code. Our solution easily integrates with any POS system, including self-checkout. A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized using our patented ViPR® technology, its UPC information is sent directly through an Ethernet connection to the POS. The cashier verifies the items that were found under the basket and continues to close the transaction. The item can remain under the basket, and with LaneHawk BOB, you are assured to get paid for it.

How it works

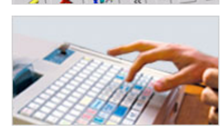
1. **Detect and recognize**
   Proprietary image recognition technology sees and identifies BOB items.



2. **Send info to POS**
   The UPC codes of the recognized items are sent via Ethernet to the POS and automatically added to the transaction.



3. **Add to transaction**
   The cashier accepts or clears the detected items to complete the sale.



Why recognizing the item is paramount

LaneHawk recognizes the specific item under the cart. It can tell the difference between a 12-pack of Coke and a 12-pack of Pepsi. It can recognize the item and ring it up without having to see the barcode. Why is item-level recognition the most important component of a BOB loss system? Because if you recognize the item, you can put that item into the current POS transaction and stop the transaction until the cashier accepts or scans in the item that was identified. Item-level recognition and a tight integration into the POS system results in a BOB loss system that eliminates almost all BOB loss. Systems which don't recognize the item and add it to the transaction are much easier for a cashier to get around.

Timeline of recognition

- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods
- Mid-late 1990s: sliding window approaches
- Late 1990s: feature-based methods
- Early 2000s – present : parts-and-shape models

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

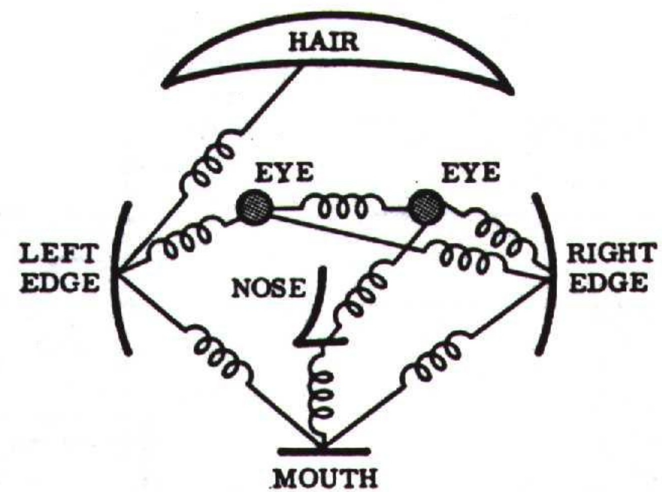


Figure from [Fischler & Elschlager 73]

Representing categories: Parts and Structure

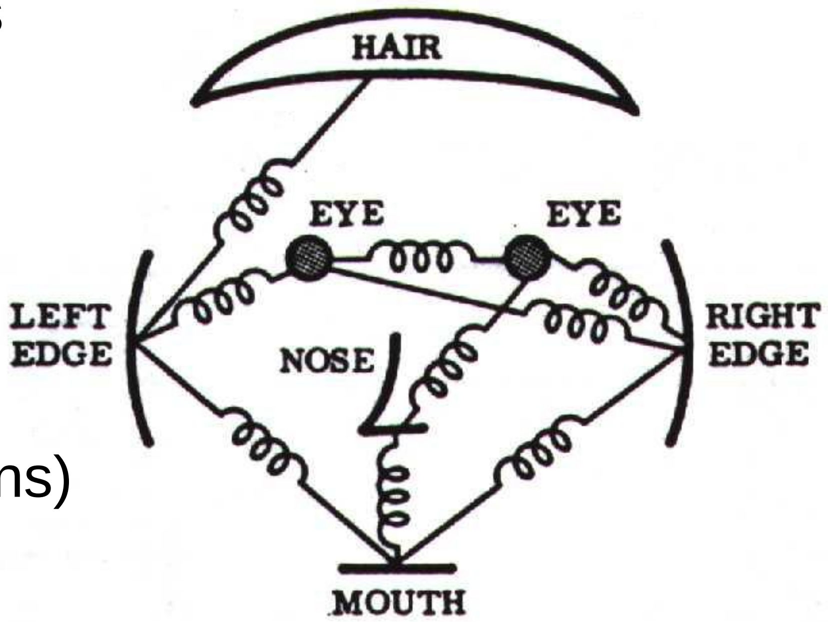
Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

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 next week

Timeline of recognition

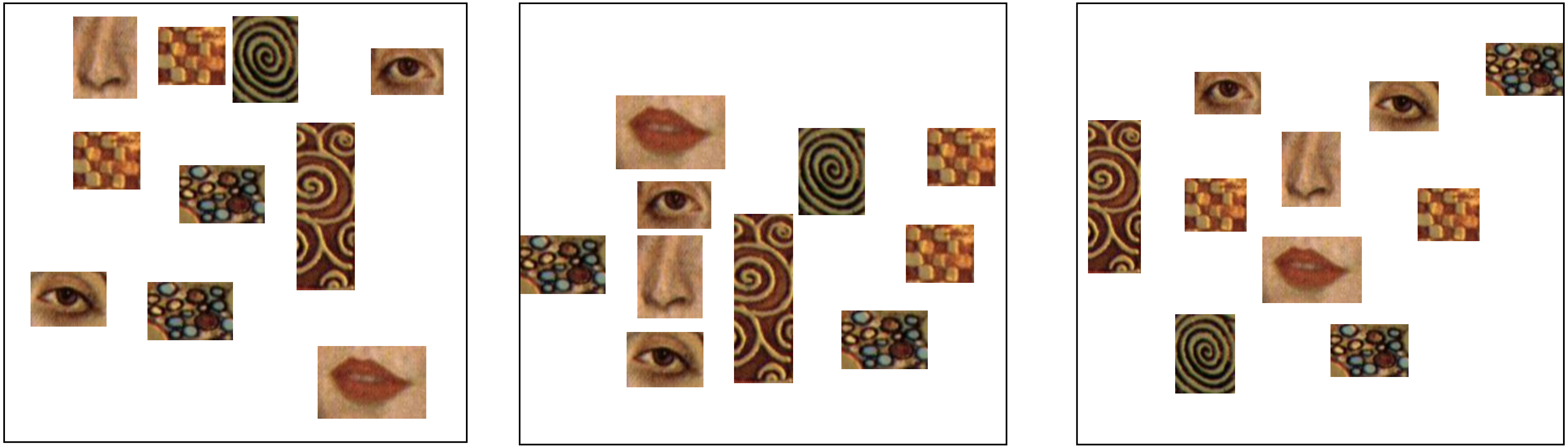
- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods
- Mid-late 1990s: sliding window approaches
- Late 1990s: feature-based methods
- Early 2000s – present : parts-and-shape models
- 2003 – present: bags of features

Bag-of-features models

Object → **Bag of 'words'**

Objects as texture

- All of these are treated as being the same



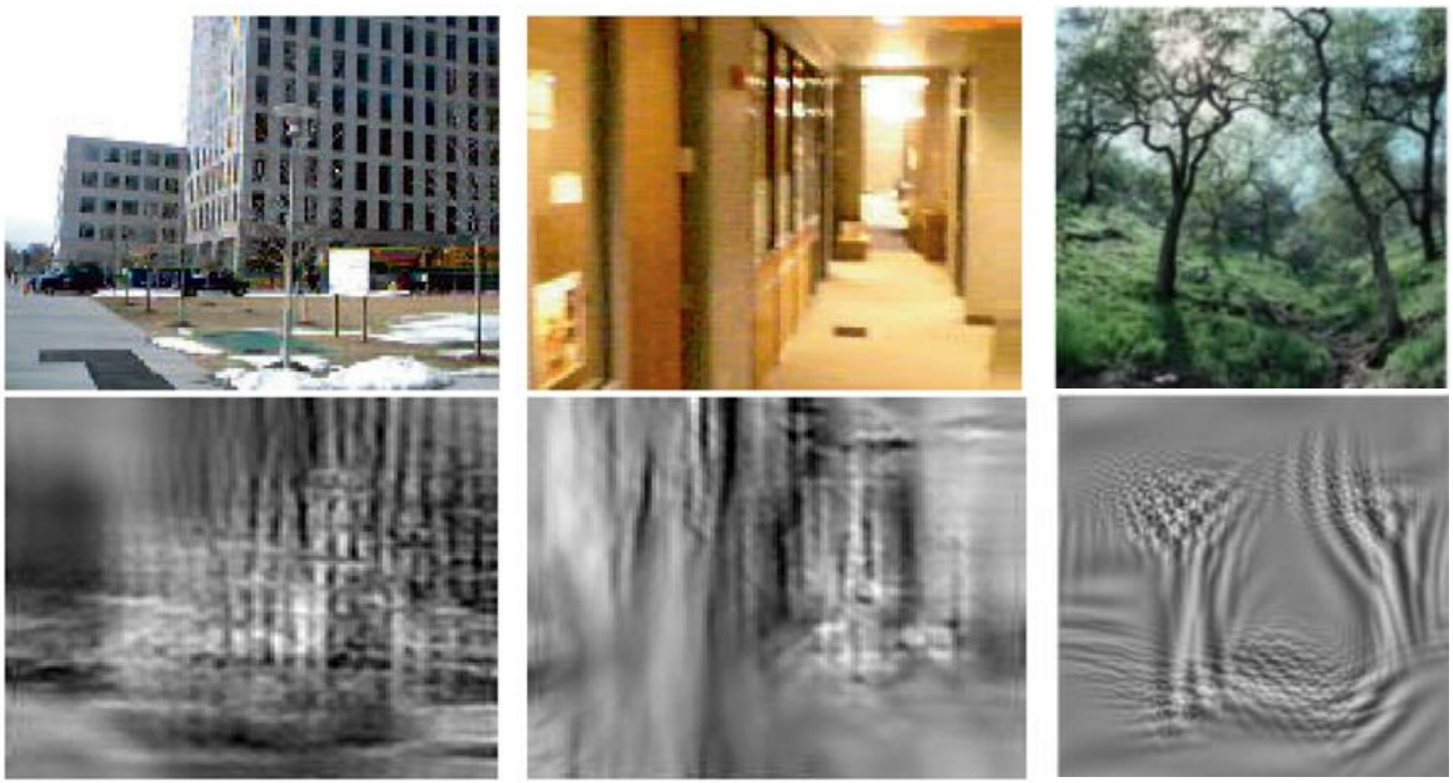
- No distinction between foreground and background: scene recognition?

Timeline of recognition

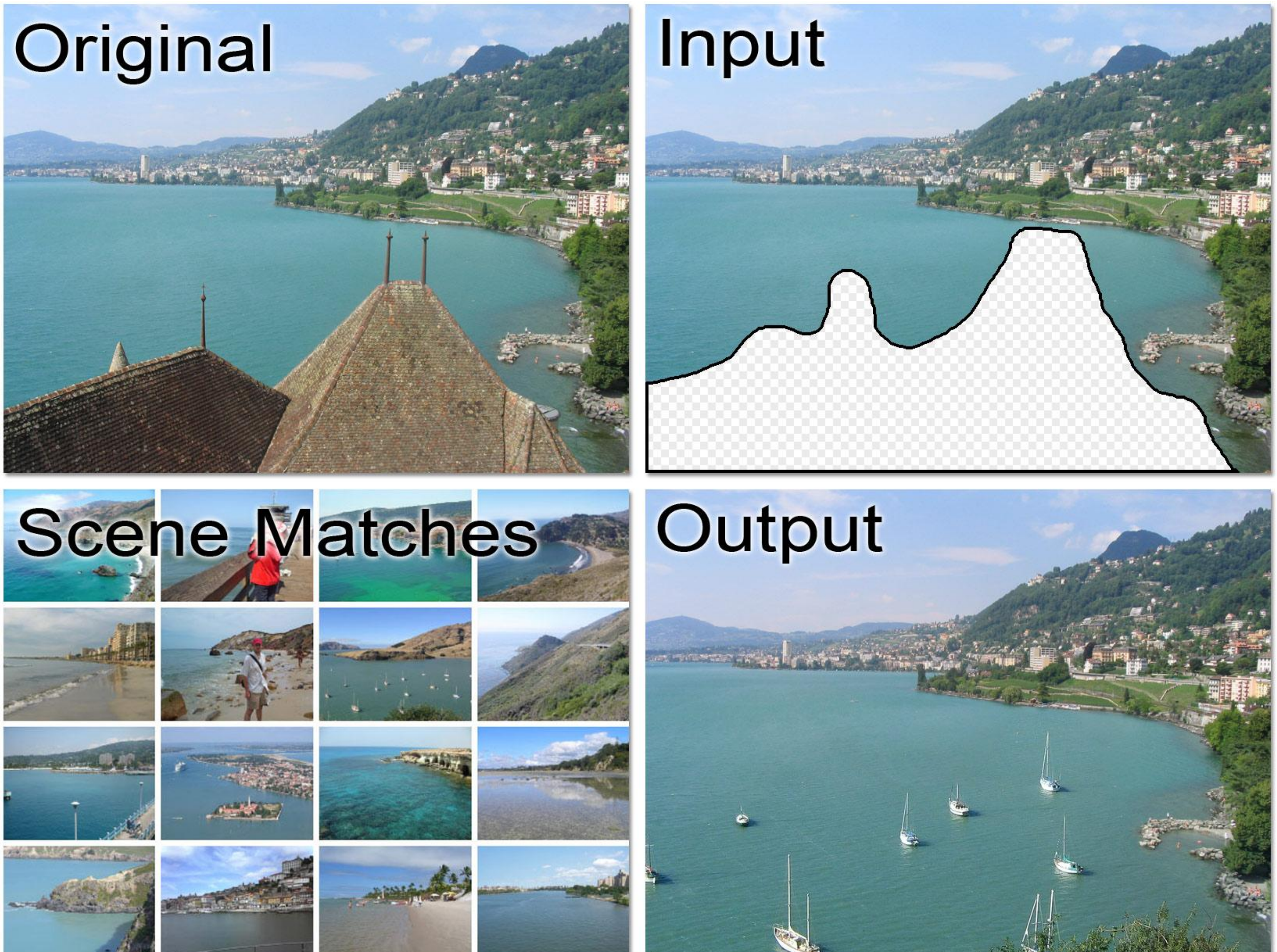
- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods
- Mid-late 1990s: sliding window approaches
- Late 1990s: feature-based methods
- Early 2000s – present : parts-and-shape models
- 2003 – present: bags of features
- Present trends: combination of local and global methods, modeling context, integrating recognition and segmentation

Global models?

- The “gist” of a scene: Oliva & Torralba (2001)



J. Hays and A. Efros, [Scene Completion using Millions of Photographs](), SIGGRAPH 2007



Object Recognition by Scene Alignment

Bryan C. Russell, Antonio Torralba, Ce Liu, Rob Fergus, William T. Freeman

NIPS 2007

Input image

Goal: Recognize objects embedded in a scene

Nearest neighbors from 15,691 images

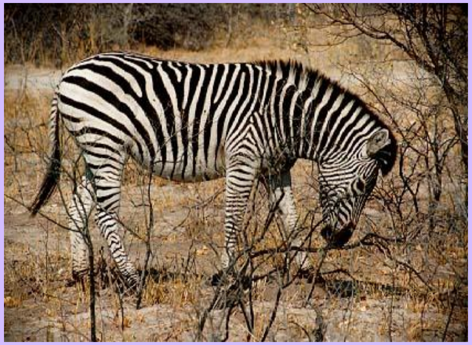
Cluster images using object labels

Output image with object labels transferred

Timeline of recognition

- 1965-late 1980s: alignment, geometric primitives
- Early 1990s: invariants, appearance-based methods
- Mid-late 1990s: sliding window approaches
- Late 1990s: feature-based methods
- Early 2000s – present : parts-and-shape models
- 2003 – present: bags of features
- Present trends: combination of local and global methods, modeling context, integrating recognition and segmentation

Object categorization: the statistical viewpoint



$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

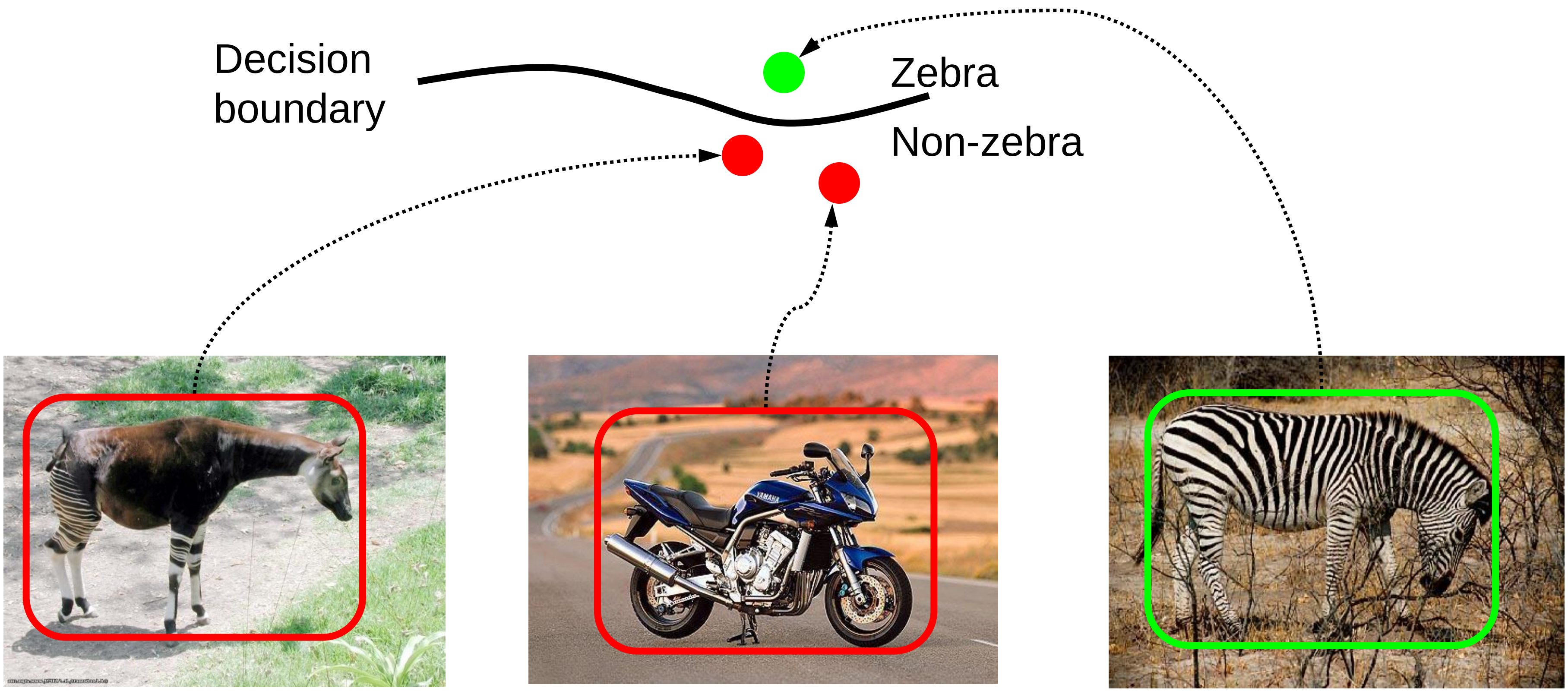
Object categorization: the statistical viewpoint

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

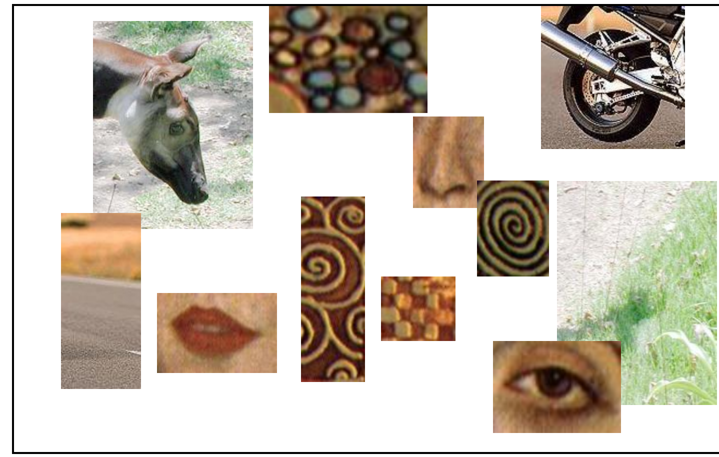
Discriminative

- Direct modeling of $$\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$$



Generative

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$



	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

Three main issues

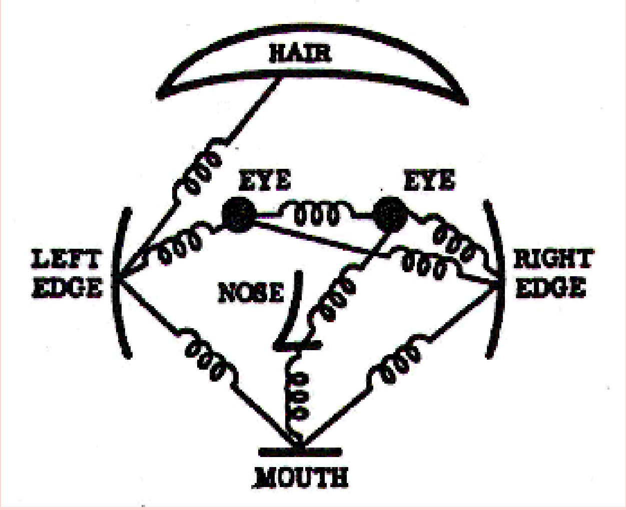
- Representation
  – How to represent an object category

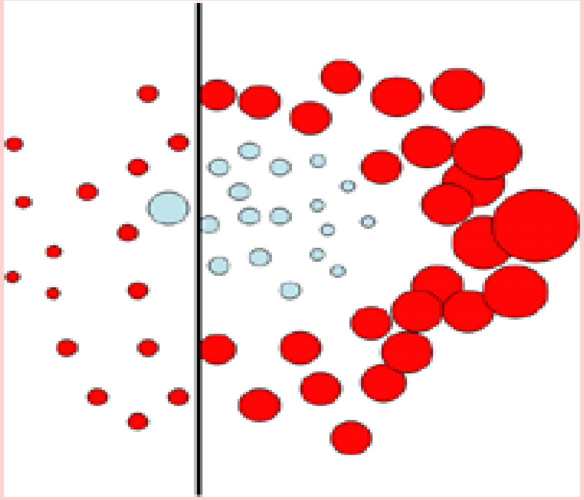
- Learning
  – How to form the classifier, given training data

- Recognition
  – How the classifier is to be used on novel data

Representation

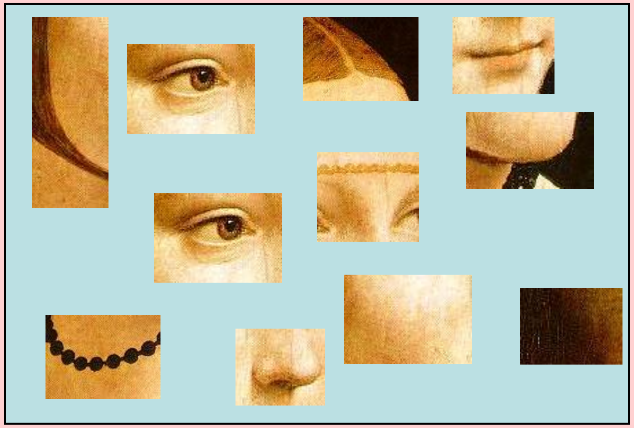
– Generative / discriminative / hybrid

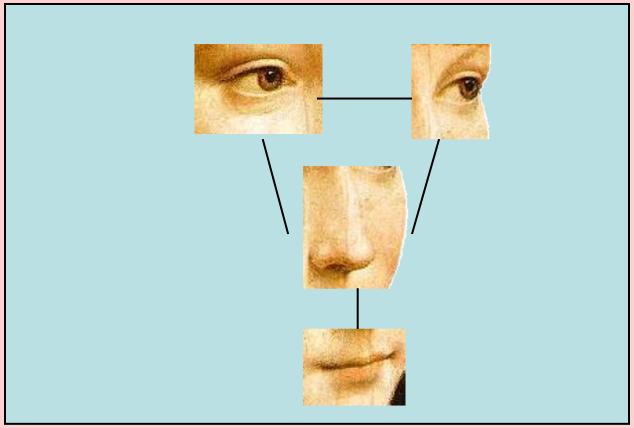




Representation

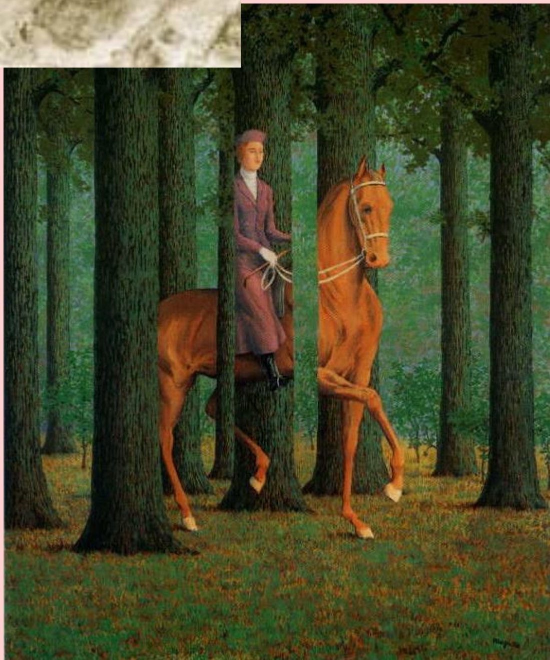
- Generative / discriminative / hybrid
- Appearance only or location and appearance





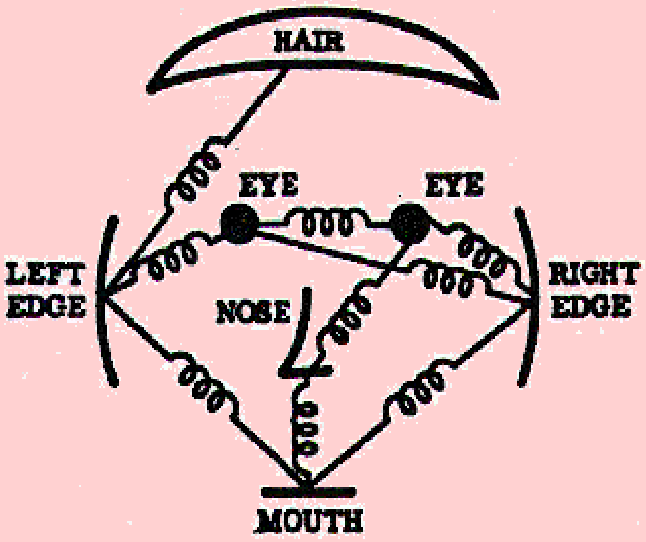
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.



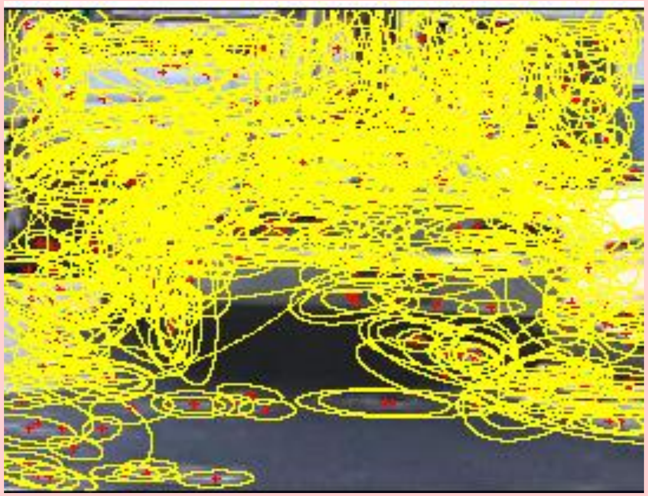
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- invariances
- Part-based or global w/sub-window



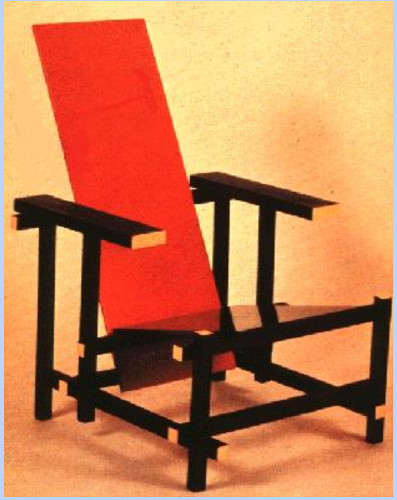
Representation

– Generative / discriminative / hybrid
– Appearance only or location and appearance
– invariances
– Parts or global w/sub-window
– Use set of features or each pixel in image



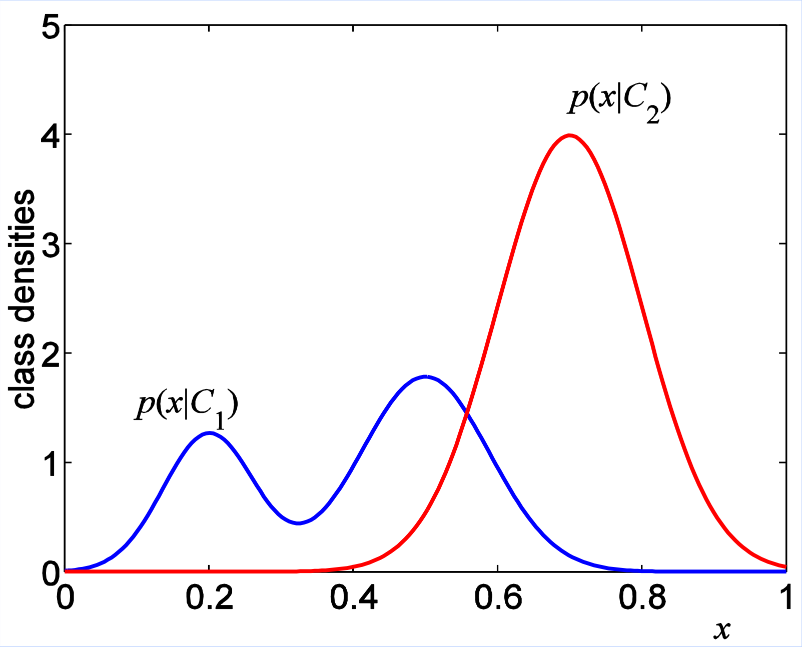
Learning

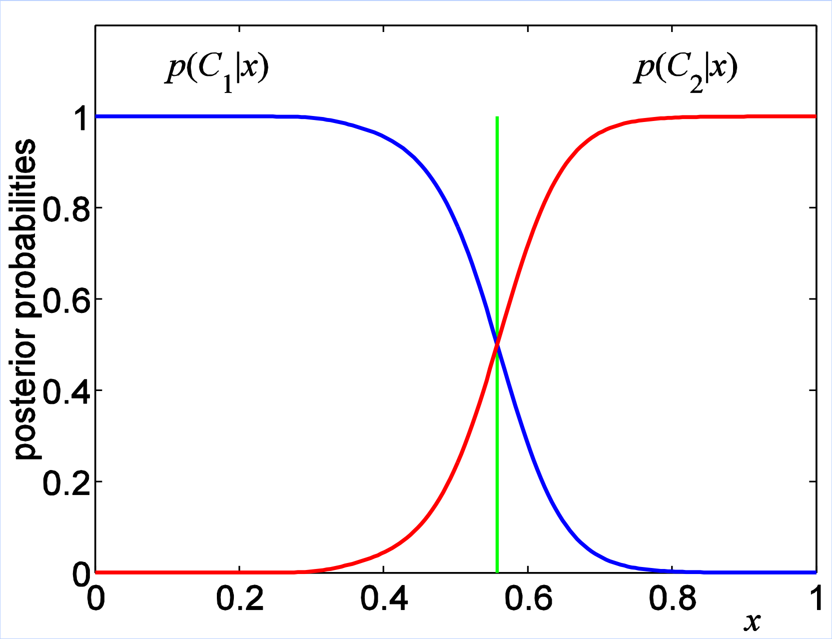
– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning



Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- Methods of training: generative vs. discriminative

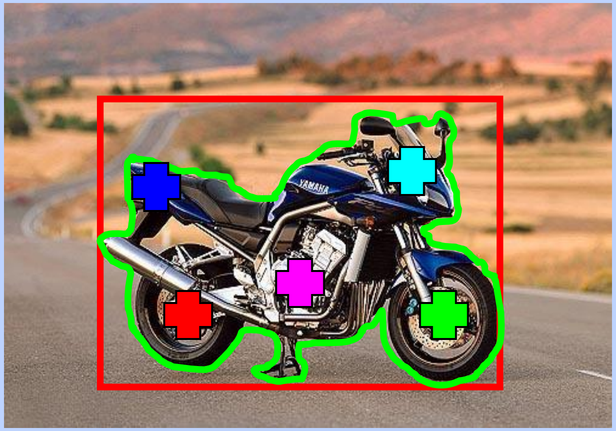




Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

Contains a motorbike



Learning

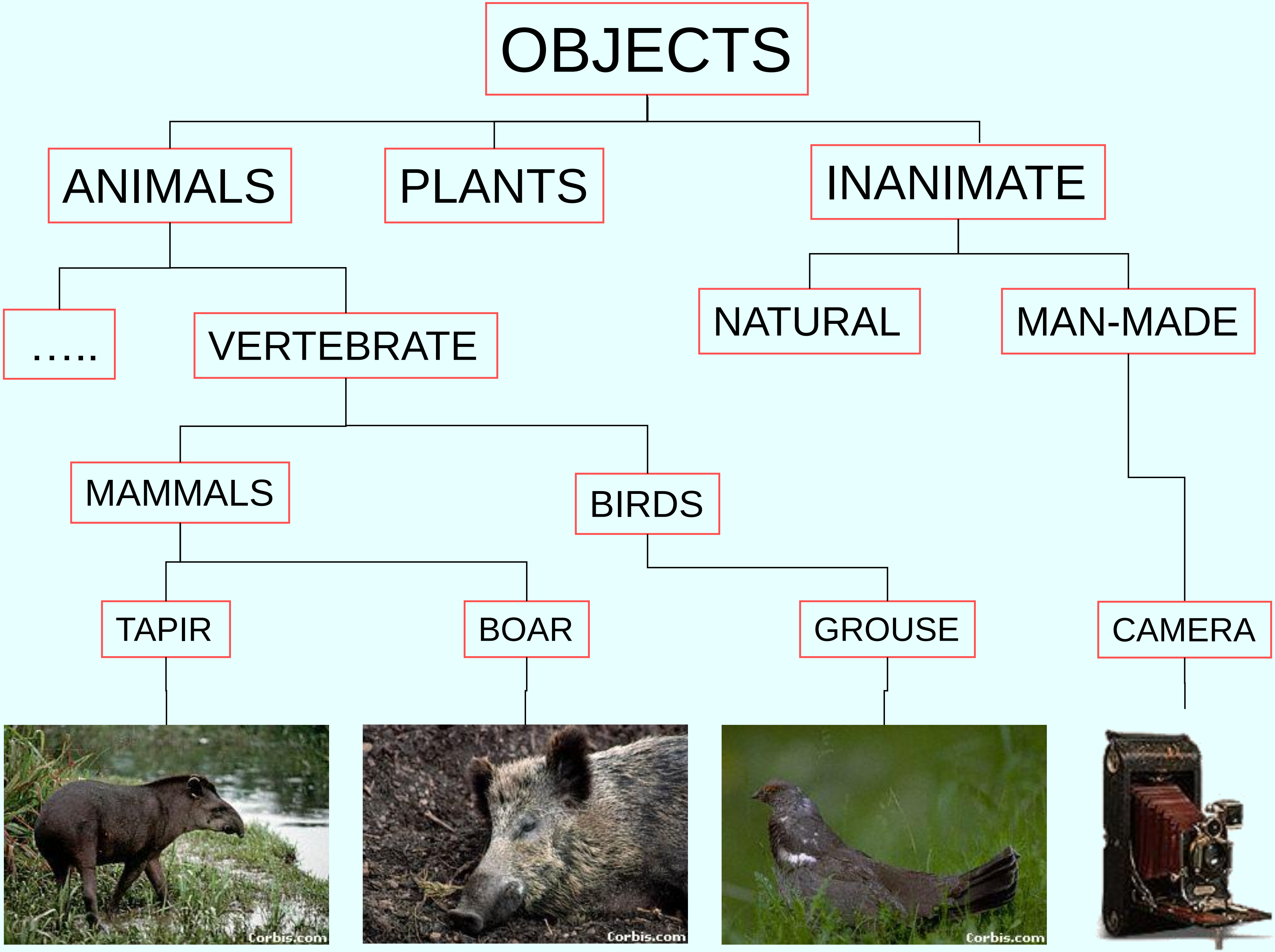
- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods Priors

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods
- Priors

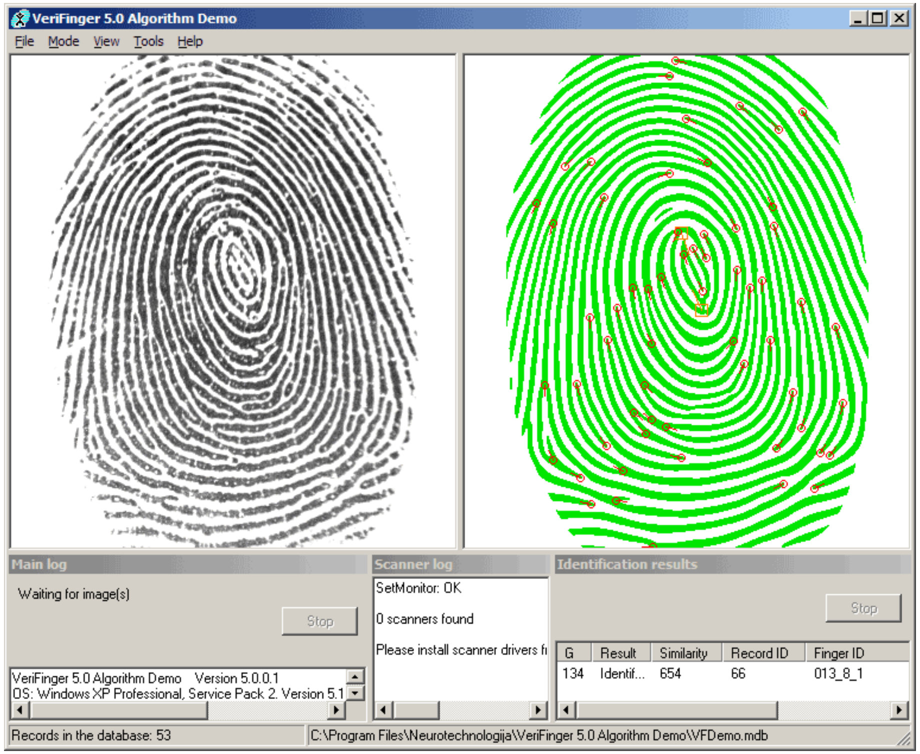

OBJECTS
ANIMALS
PLANTS
INANIMATE
.....
VERTEBRATE
NATURAL
MAN-MADE
MAMMALS
BIRDS
TAPIR
BOAR
GROUSE
CAMERA
Corbis.com
Corbis.com
Corbis.com

What “works” today

- Reading license plates, zip codes, checks

What “works” today

- Reading license plates, zip codes, checks
- Fingerprint recognition



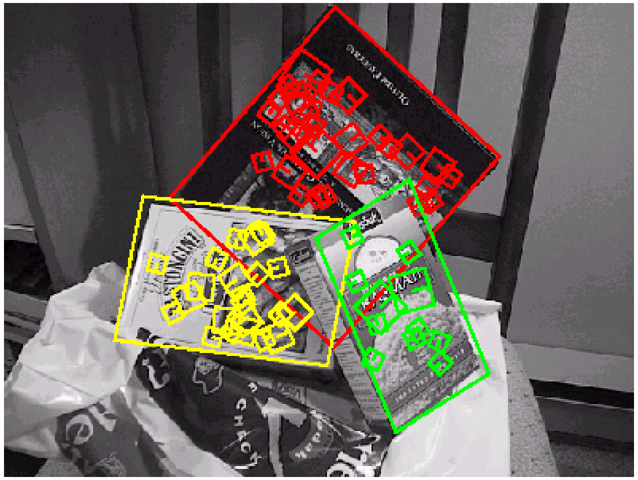
What “works” today

- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection

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

What "works" today

- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection
- Recognition of flat textured objects (CD covers, book covers, etc.)



Specific Object Recognition Application

Evolution Robotics Retail™ Our vision, your profit

Home | Products | Our Company | News & Events | CONTACT US >

Products

- LaneHawk BOB
  - Benefits
  - Store Integration
  - Demonstration Video
  - Stopping BOB Loss
- LaneHawk InCart
- LaneHawk Enterprise Manager
- ShelfHawk
- TunnelHawk
- Core Technologies

LaneHawk® BOB

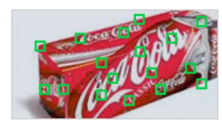
LaneHawk® BOB is a loss-prevention solution that turns bottom-of-basket (BOB) losses into profits in real time.

WATCH DEMO VIDEO >

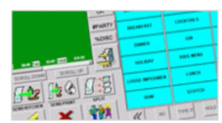
It starts with a visual scanner that recognizes items without having to read the UPC code. Our solution easily integrates with any POS system, including self-checkout. A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized using our patented ViPR® technology, its UPC information is sent directly through an Ethernet connection to the POS. The cashier verifies the items that were found under the basket and continues to close the transaction. The item can remain under the basket, and with LaneHawk BOB, you are assured to get paid for it.

How it works

1. **Detect and recognize**
   Proprietary image recognition technology sees and identifies BOB items.



2. **Send info to POS**
   The UPC codes of the recognized items are sent via Ethernet to the POS and automatically added to the transaction.



3. **Add to transaction**
   The cashier accepts or clears the detected items to complete the sale.

Why recognizing the item is paramount

LaneHawk recognizes the specific item under the cart. It can tell the difference between a 12-pack of Coke and a 12-pack of Pepsi. It can recognize the item and ring it up without having to see the barcode. Why is item-level recognition the most important component of a BOB loss system? Because if you recognize the item, you can put that item into the current POS transaction and stop the transaction until the cashier accepts or scans in the item that was identified. Item-level recognition and a tight integration into the POS system results in a BOB loss system that eliminates almost all BOB loss. Systems which don't recognize the item and add it to the transaction are much easier for a cashier to get around.