Object Recognition: Overview and History

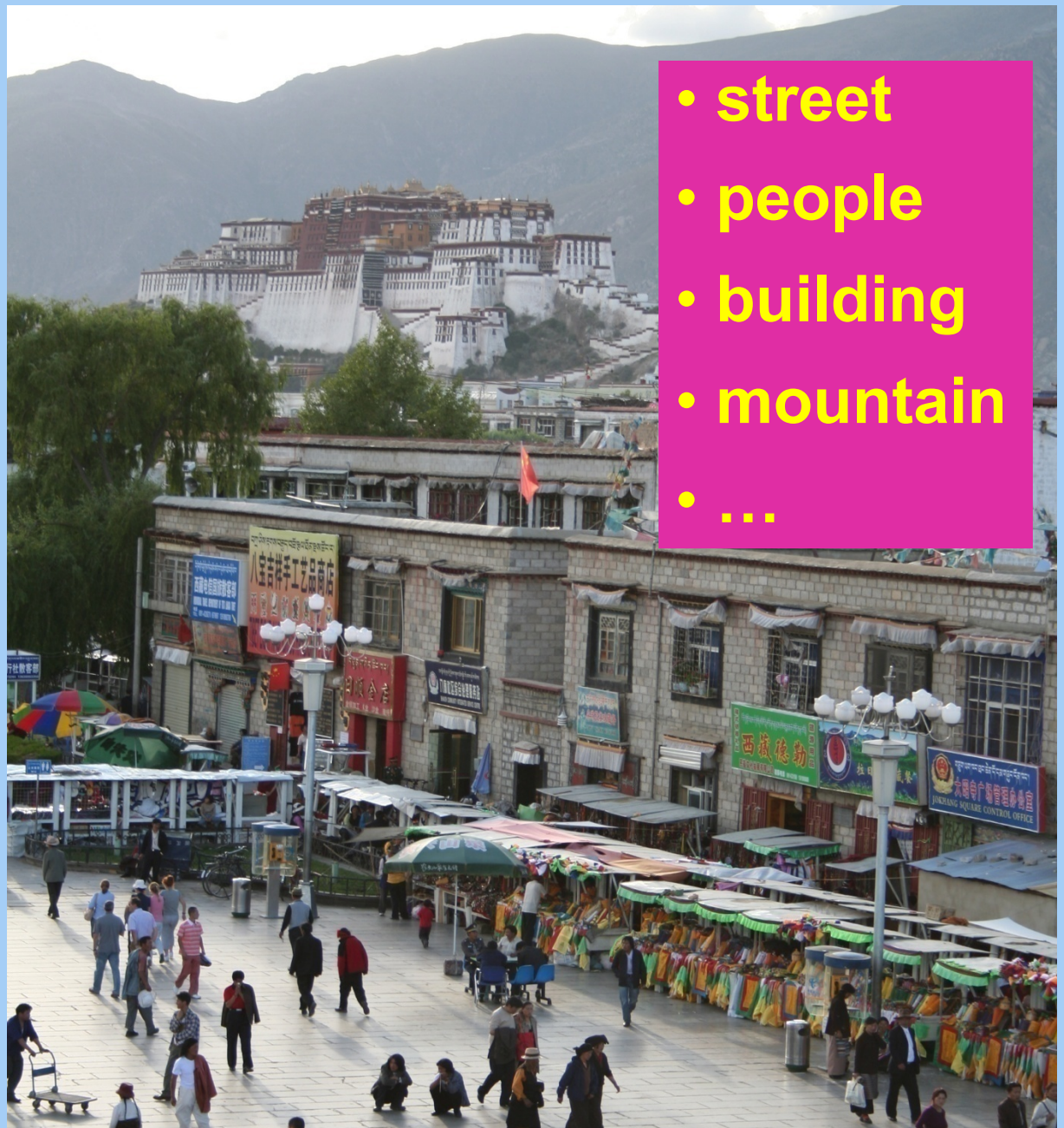
Slides adapted from Fei-Fei Li, Rob Fergus, Antonio Torralba, and Jean Ponce

Specific recognition tasks

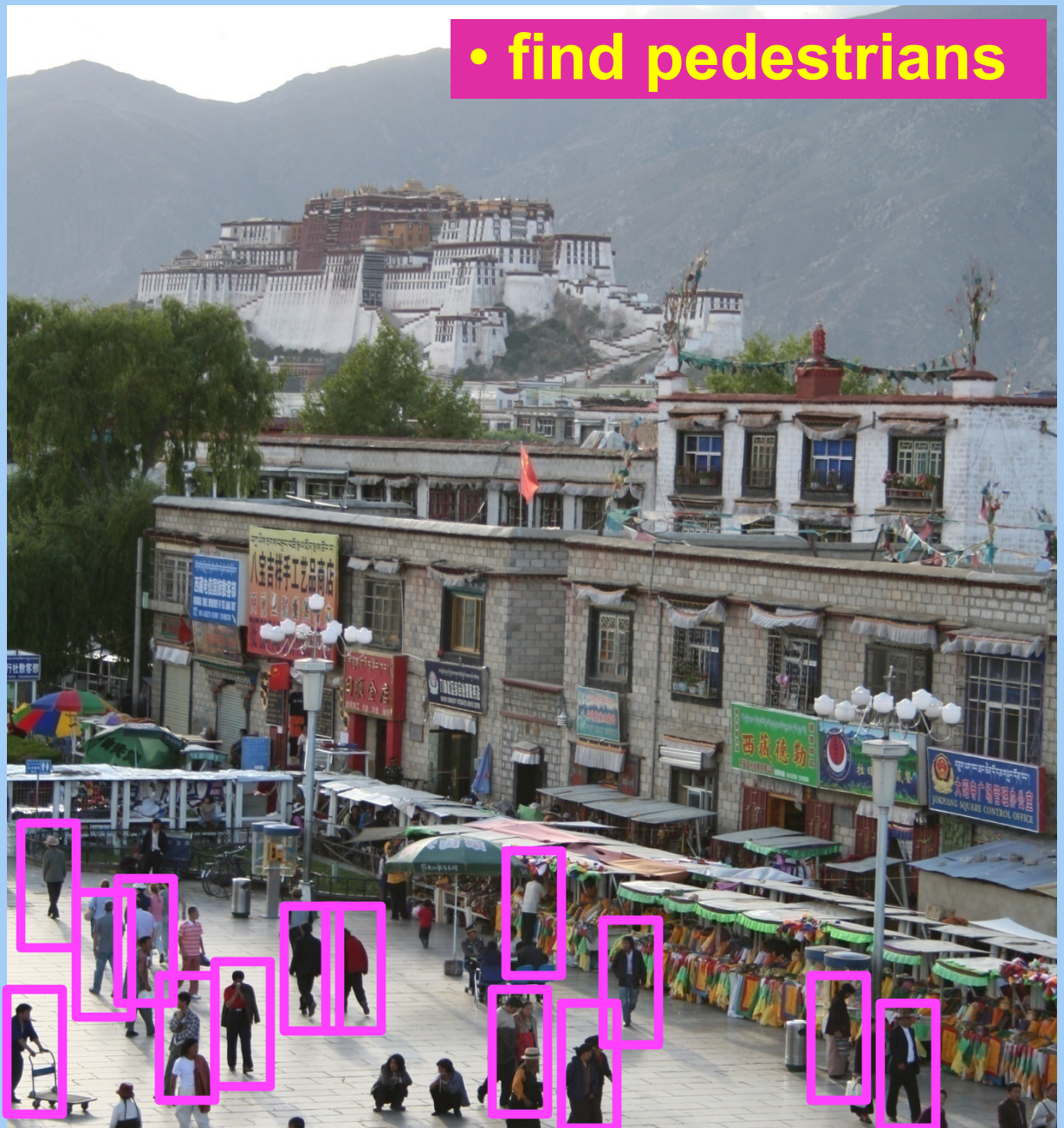
Scene categorization

- outdoor/indoor
- city/forest/factory/etc.

Image annotation/tagging



Object detection



Activity recognition

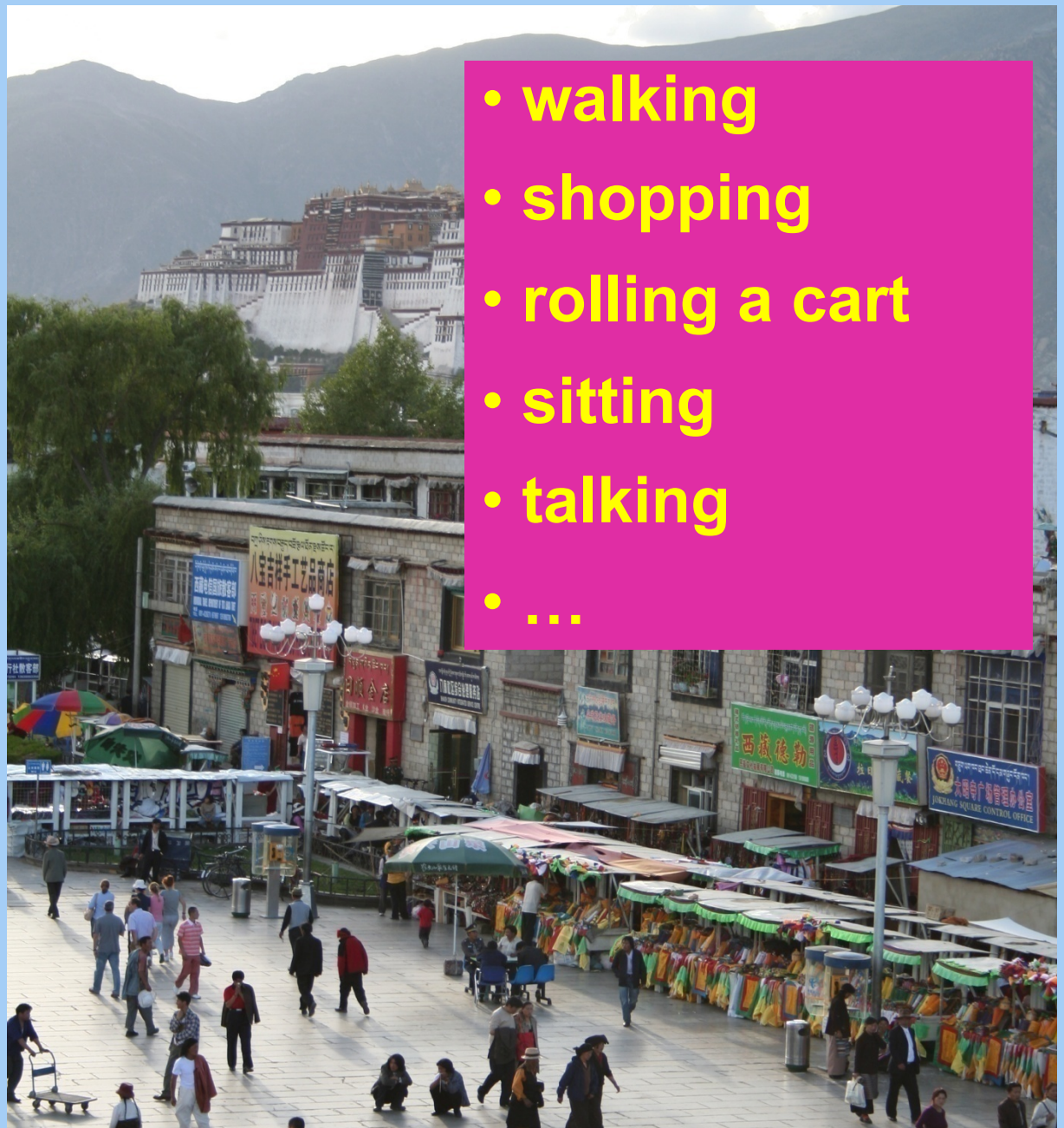
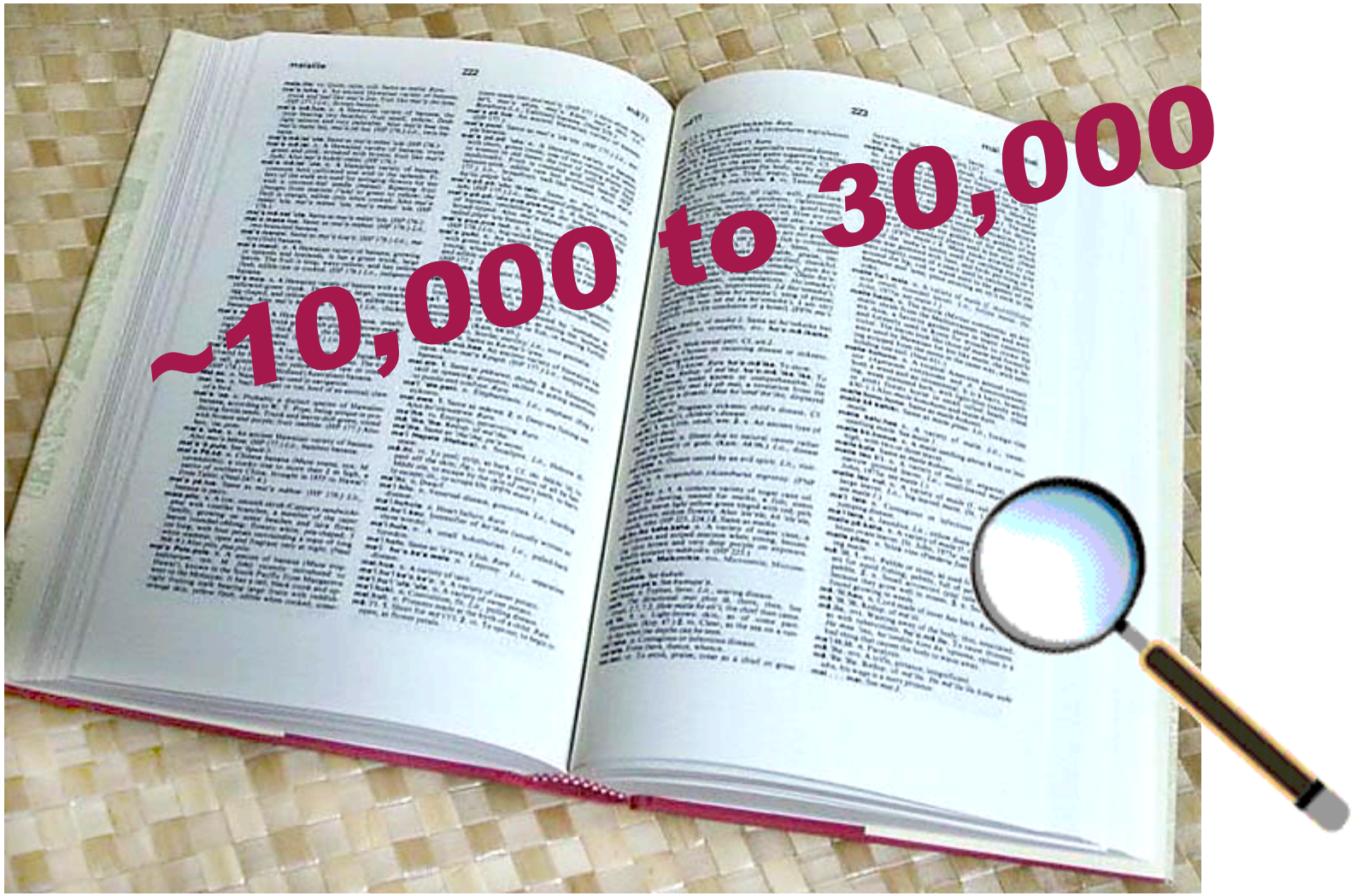


Image parsing

Image understanding?

How many visual object categories are there?



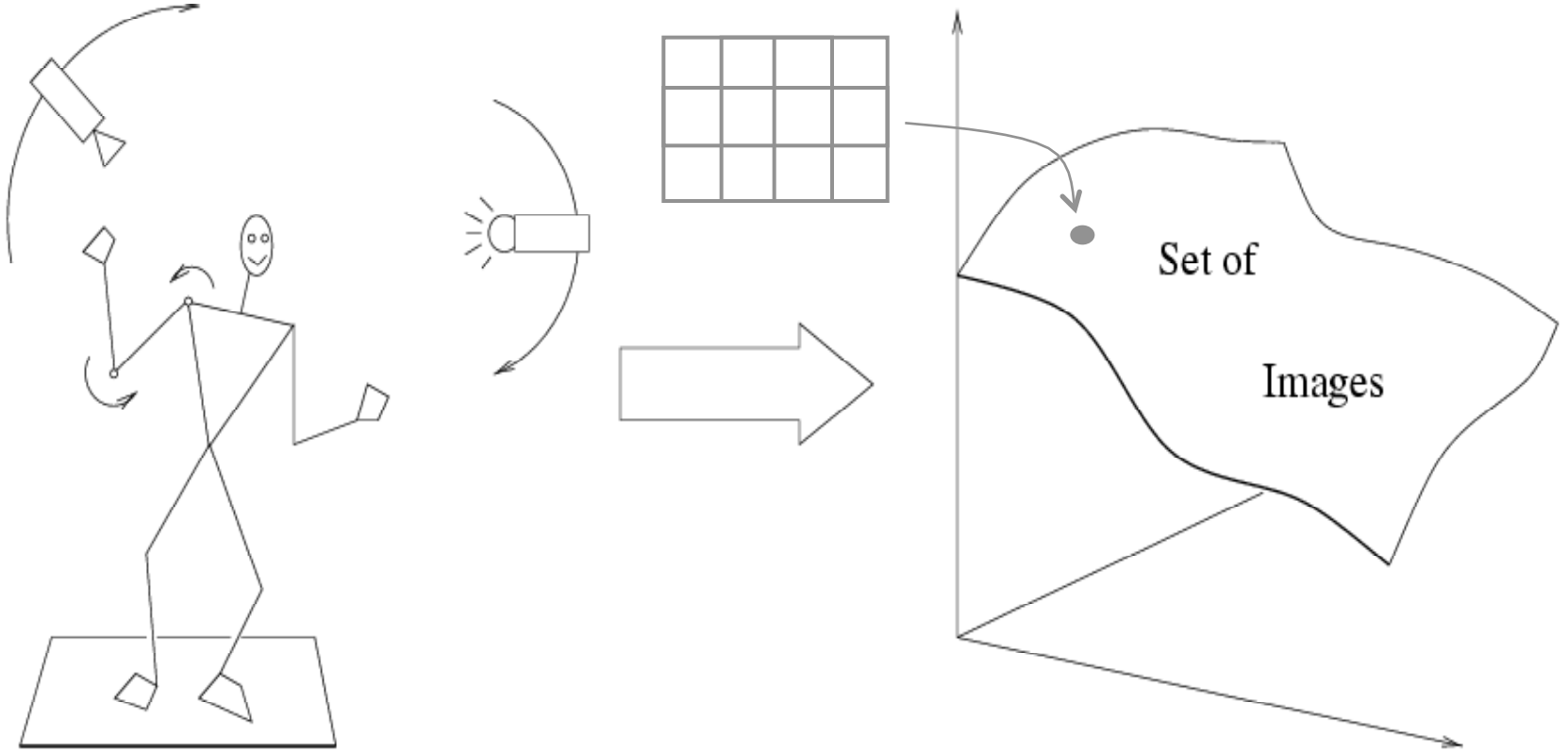
Biederman 1987


~10,000 to 30,000

OBJECTS

- ANIMALS
  - …..
  - VERTEBRATE
    - MAMMALS
      - TAPIR
      - BOAR
    - BIRDS
      - GROUSE
- PLANTS
- INANIMATE
  - NATURAL
  - MAN-MADE
    - CAMERA

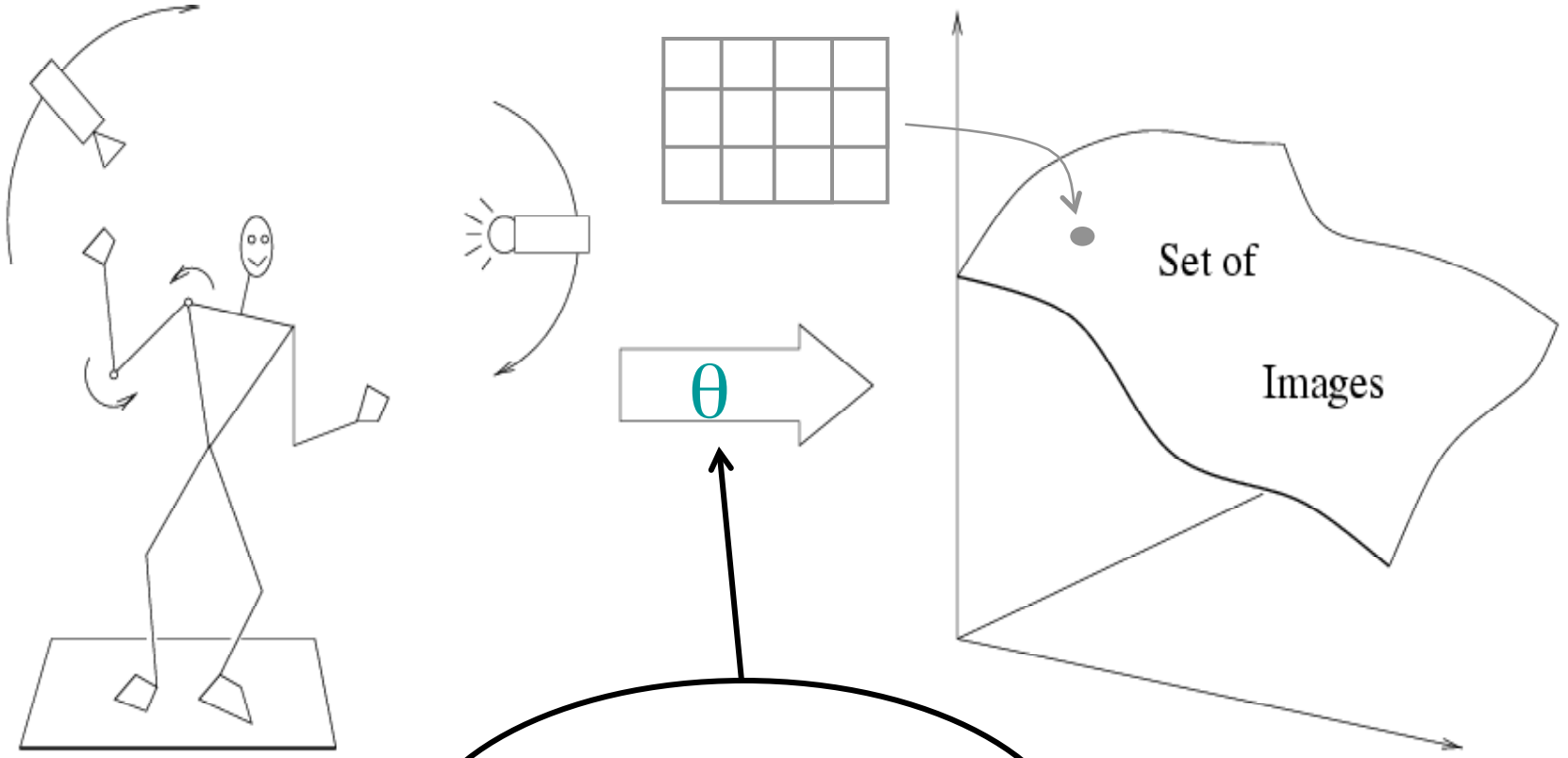
Recognition is all about modeling variability



Variability: Camera position
Illumination
Within-class variation
Background, occlusion

History of ideas in recognition

- **1960s – early 1990s: the geometric era**



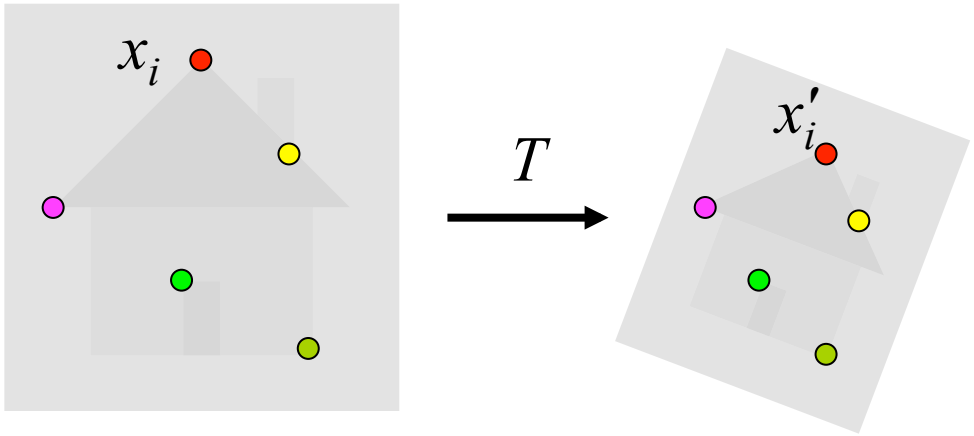
Variability: camera position **Alignment**

Shape: assumed known

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

Recall: Alignment

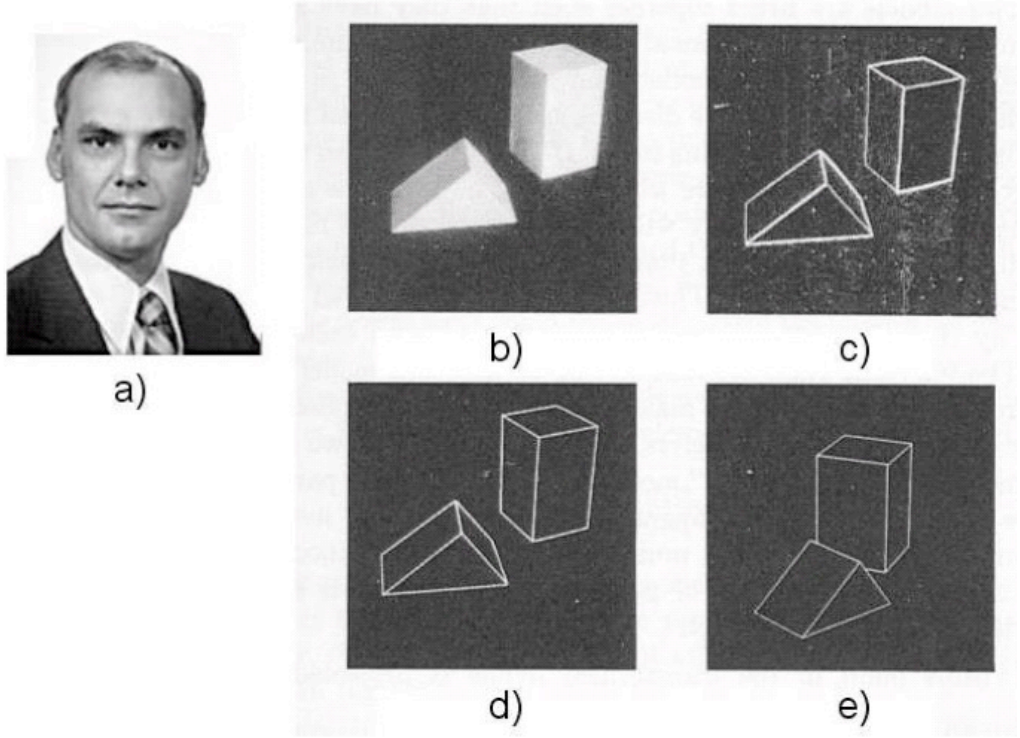
- Alignment: fitting a model to a transformation between pairs of features (*matches*) in two images



Find transformation T that minimizes

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

Recognition as an alignment problem: Block world

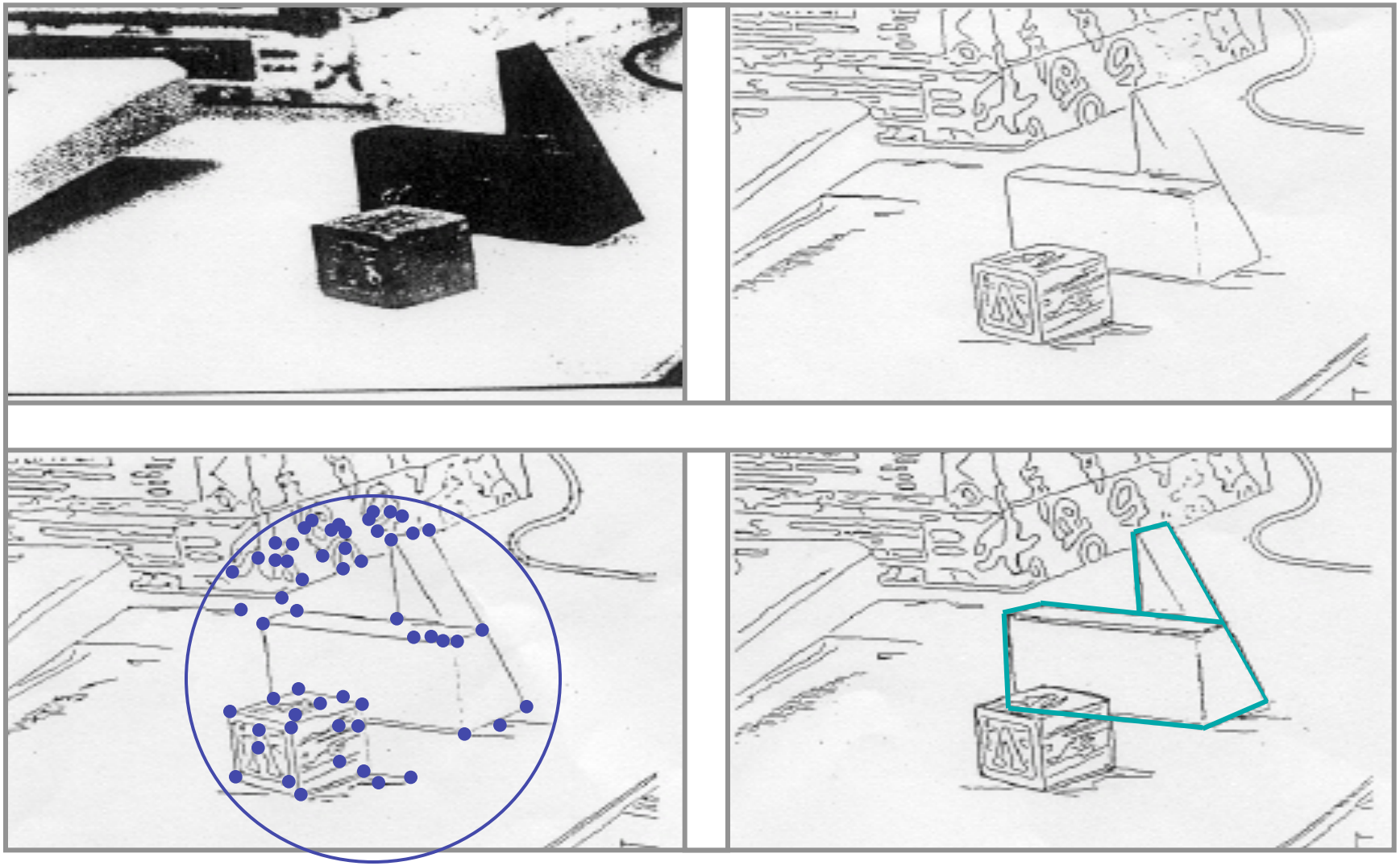


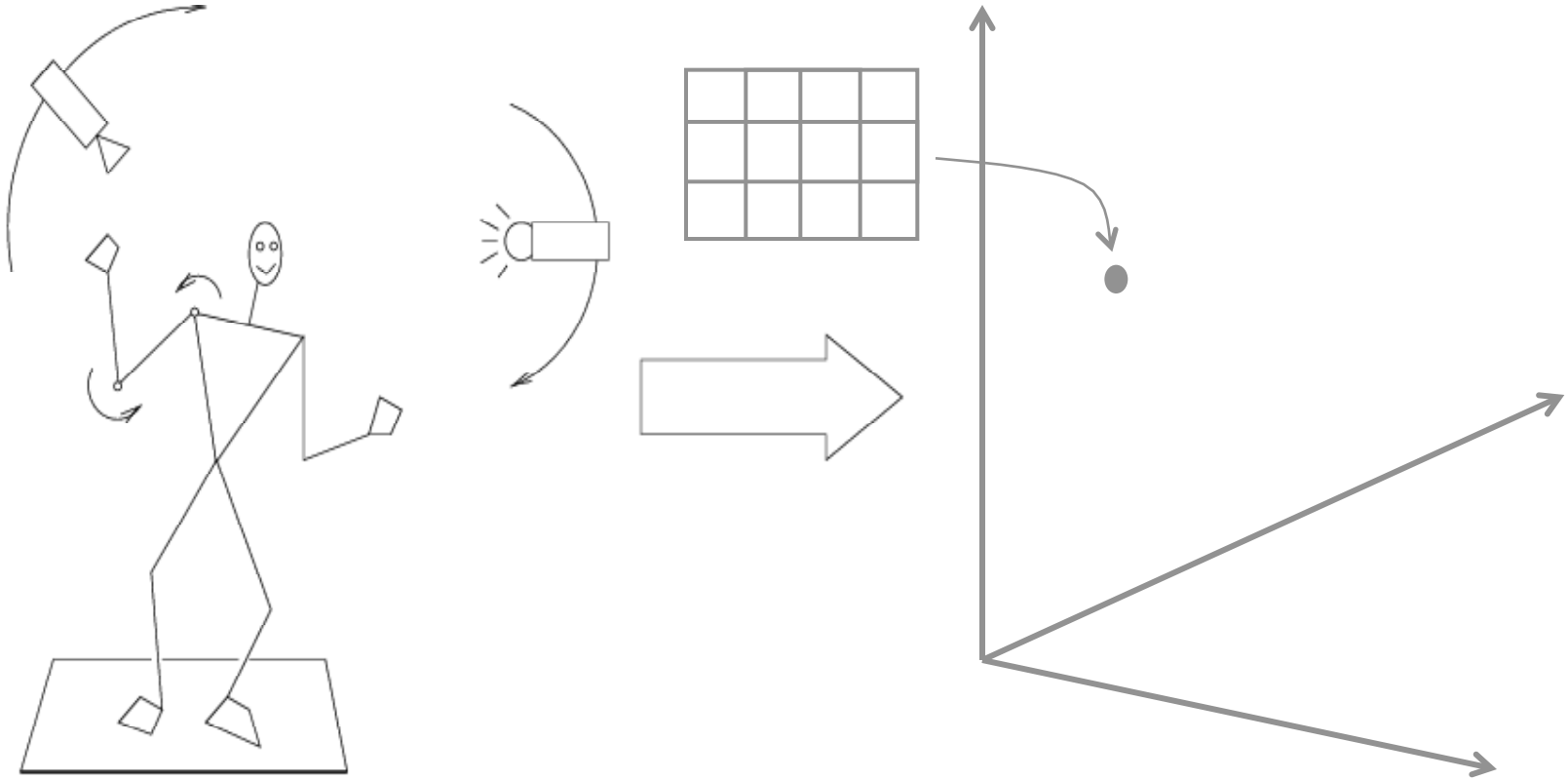
L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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.)

J. Mundy, Object Recognition in the Geometric Era: a Retrospective, 2006

Alignment: Huttenlocher & Ullman (1987)





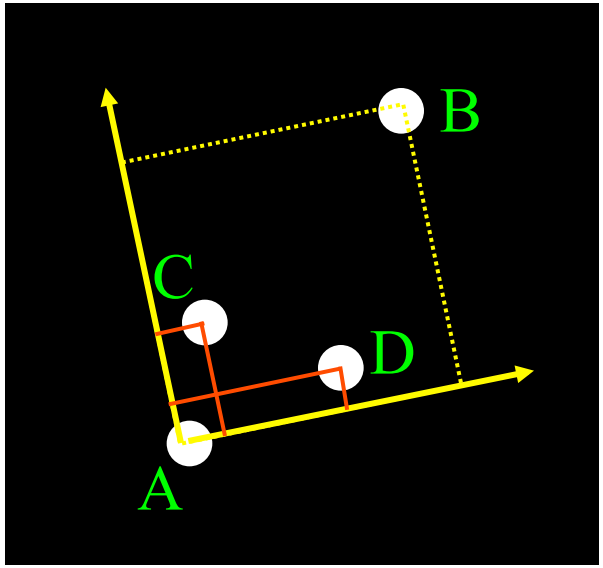
~~Variability~~

Invariance to:

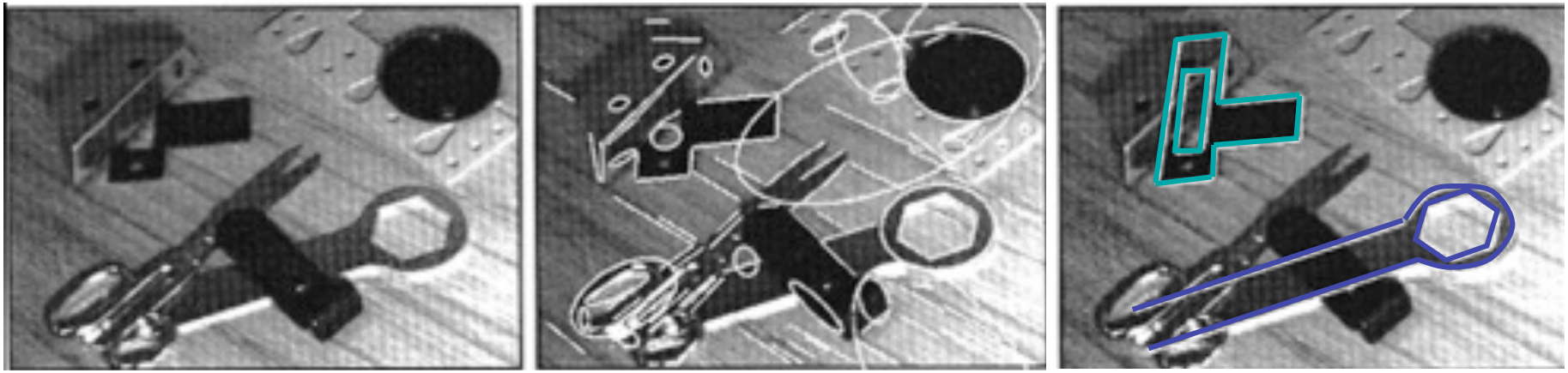
Camera position
Illumination
Etc.

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

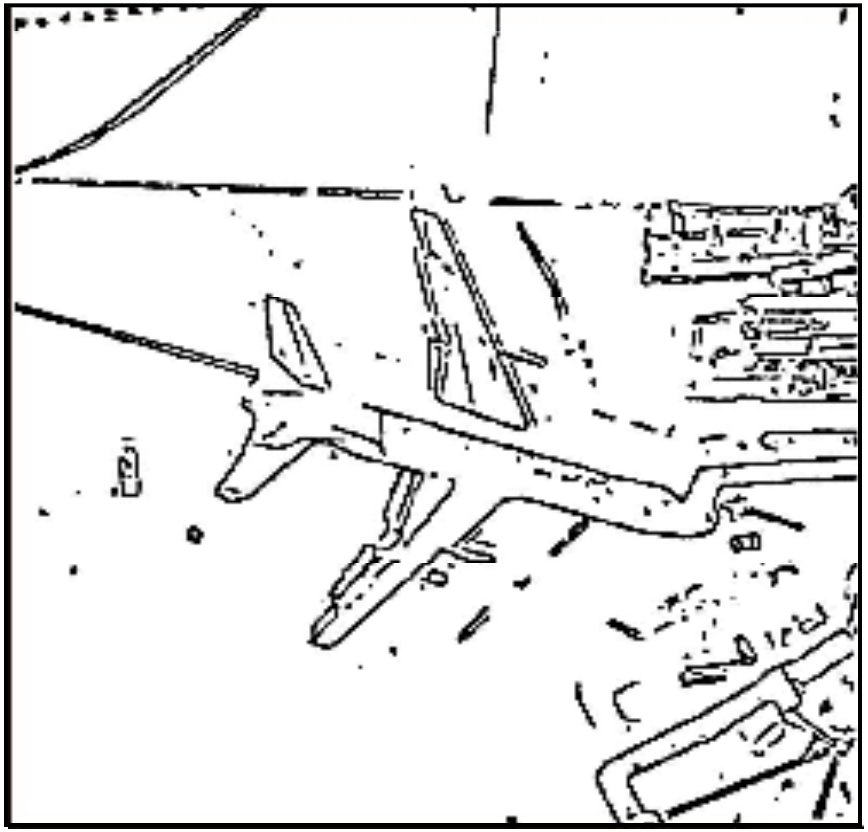


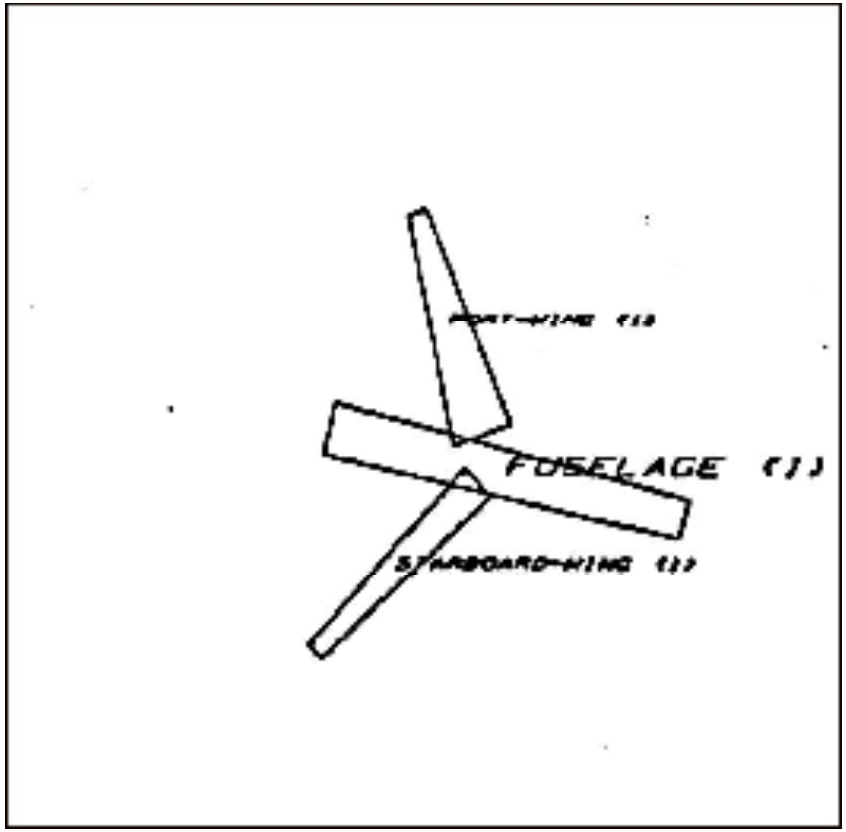
Projective invariants (Rothwell et al., 1992):



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

From object instances to object categories





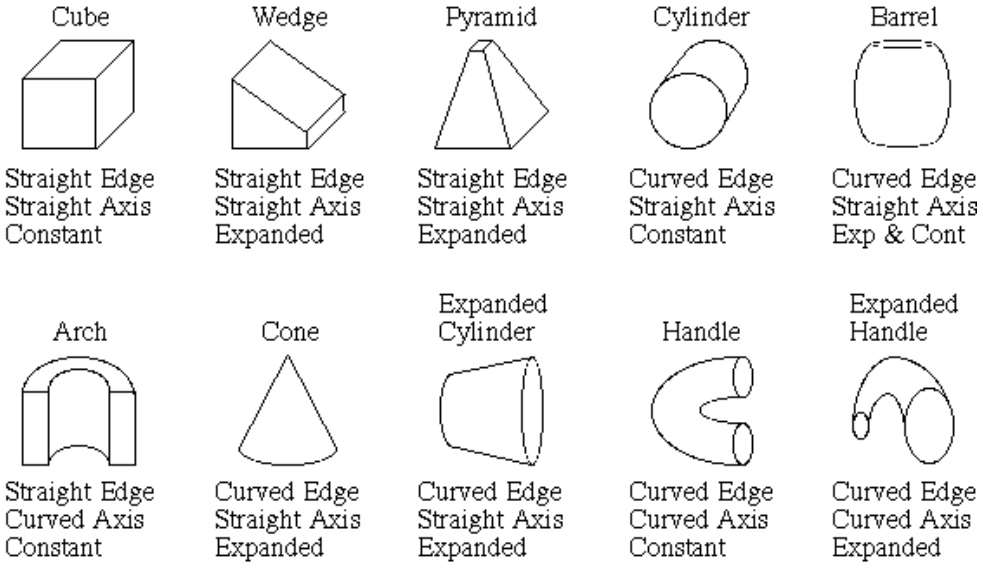
ACRONYM (Brooks and Binford, 1981)

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

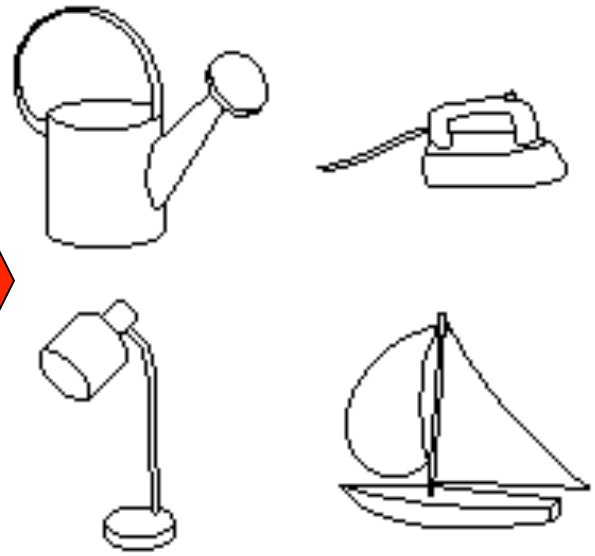
Recognition by components

Biederman (1987)

Primitives (*geons*)

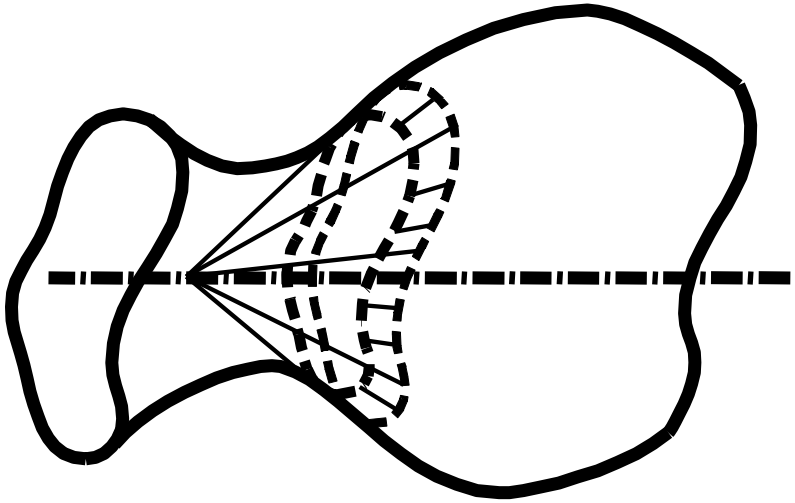


Objects

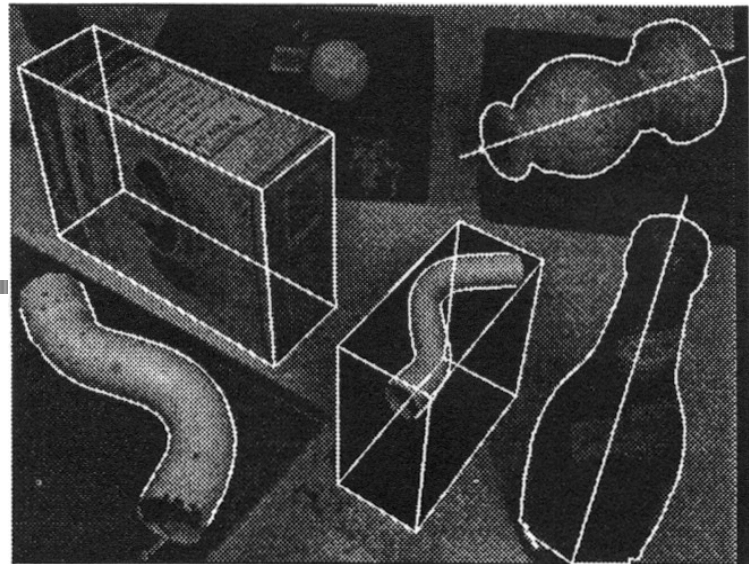


http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

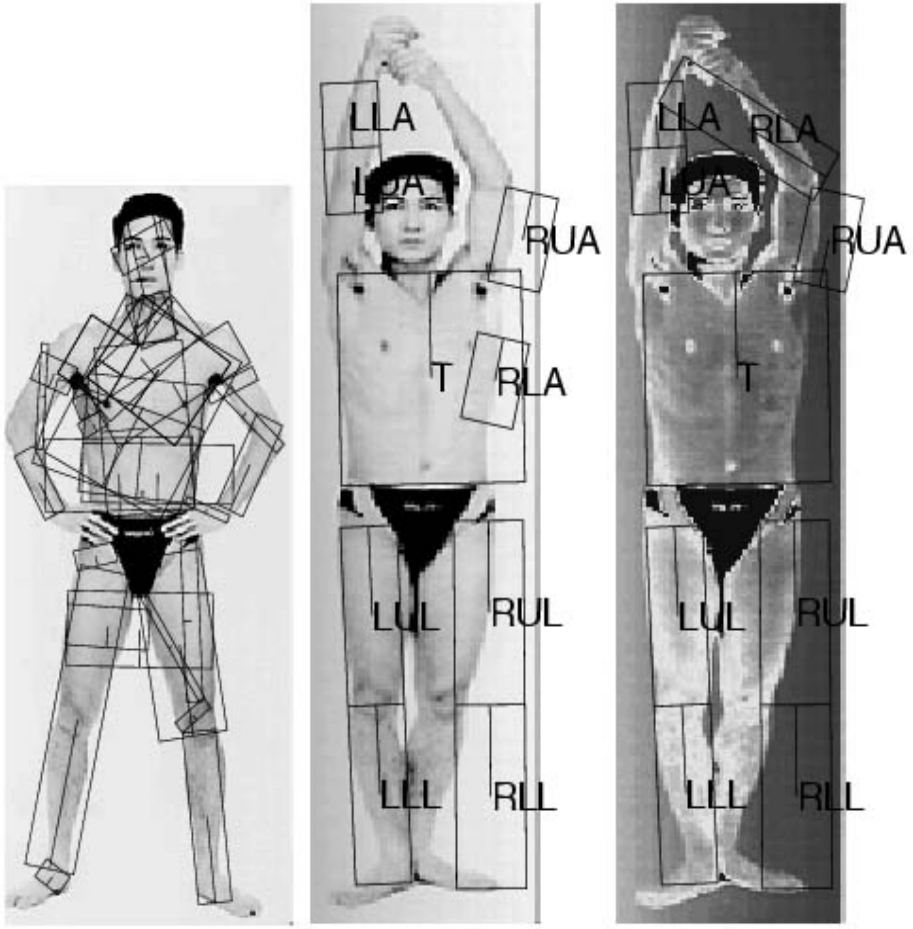
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



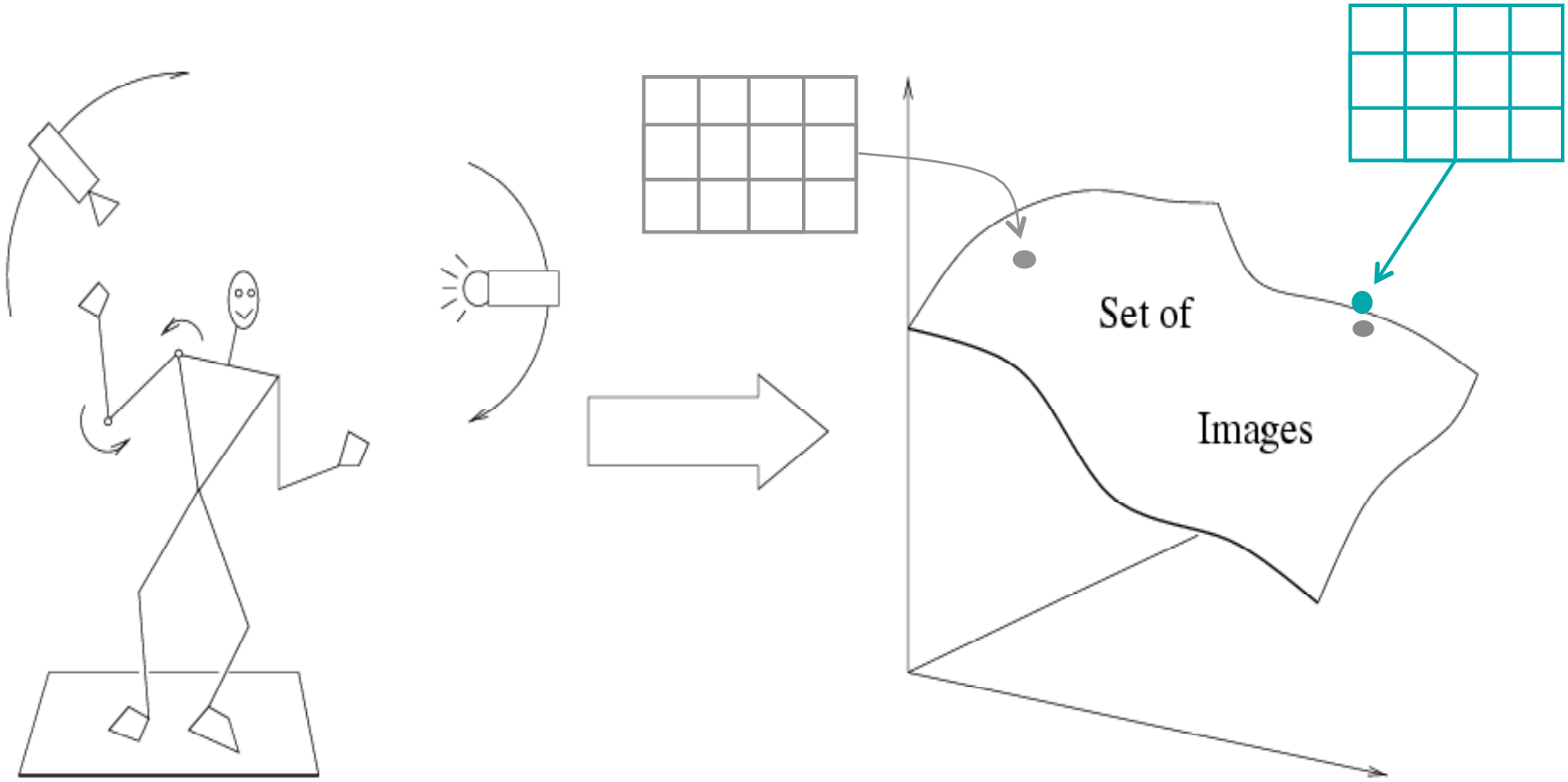
Zisserman et al. (1995)



Forsyth (2000)

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- **1990s: appearance-based models**

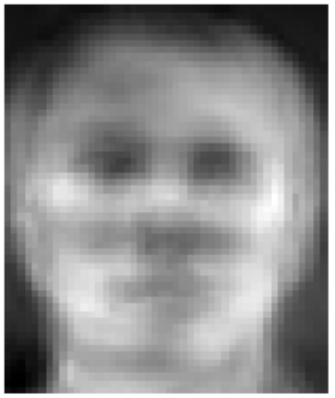


Empirical models of image variability

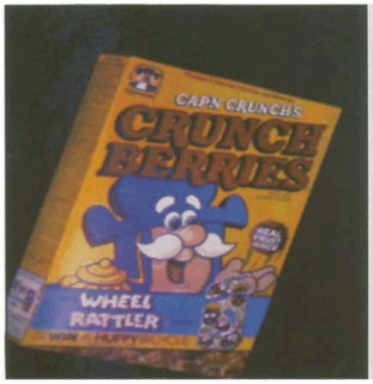
Appearance-based techniques

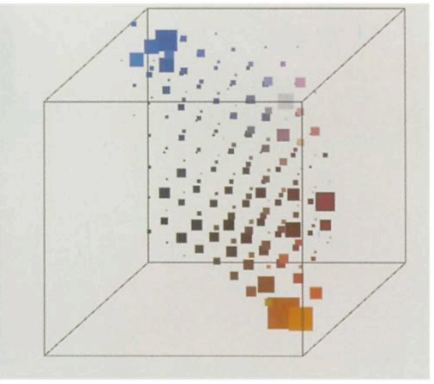
Turk & Pentland (1991); Murase & Nayar (1995); etc.

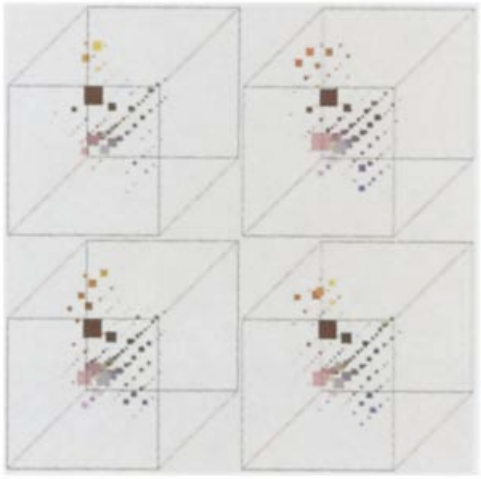
Eigenfaces (Turk & Pentland, 1991)



Color Histograms

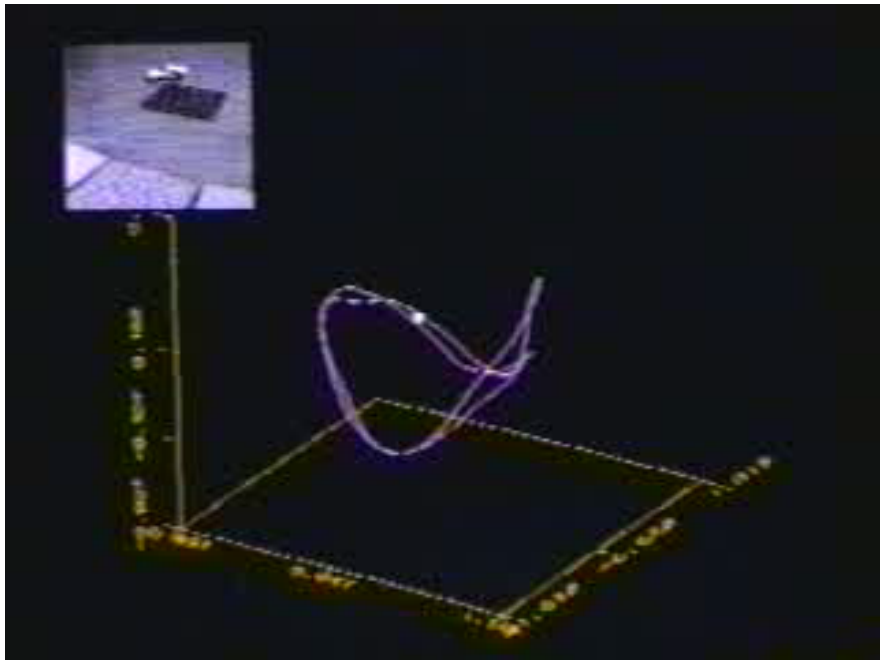






Swain and Ballard, Color Indexing, IJCV 1991.

Appearance manifolds



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

Limitations of global appearance models

- Requires global registration of patterns
- Not robust to clutter, occlusion, geometric transformations

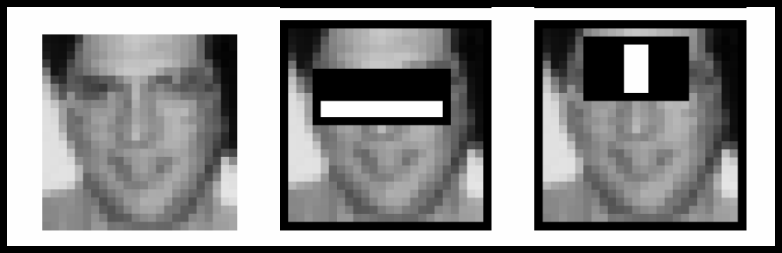
History of ideas in recognition

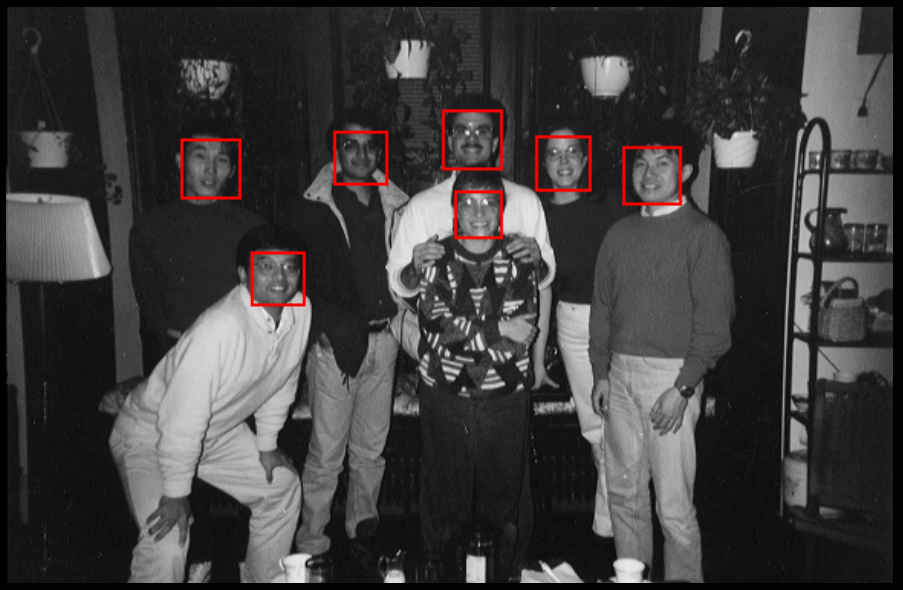
- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- **1990s – present: sliding window approaches**

Sliding window approaches

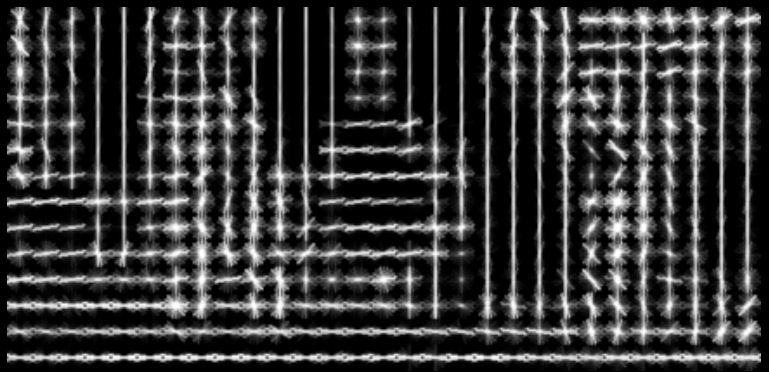
Sliding window approaches

- Viola and Jones, 2000

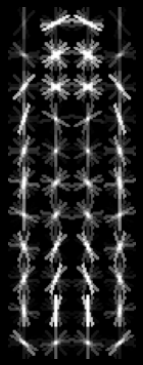




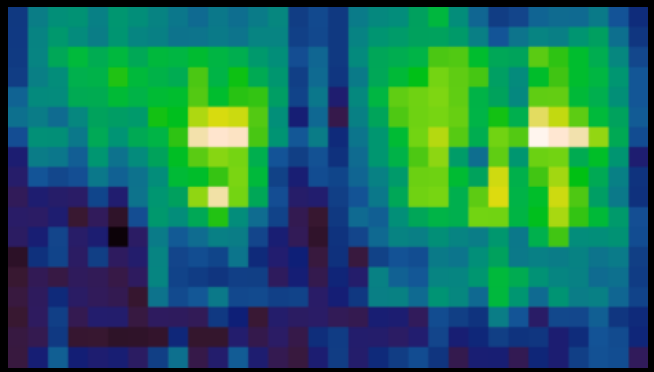
- Dalal and Triggs, 2005



HOG feature map



Template



Detector response map

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- 1990s – present: sliding window approaches
- **Late 1990s: local features**

Local features for object instance recognition

D. Lowe (1999, 2004)

Large-scale image search

Combining local features, indexing, and spatial constraints

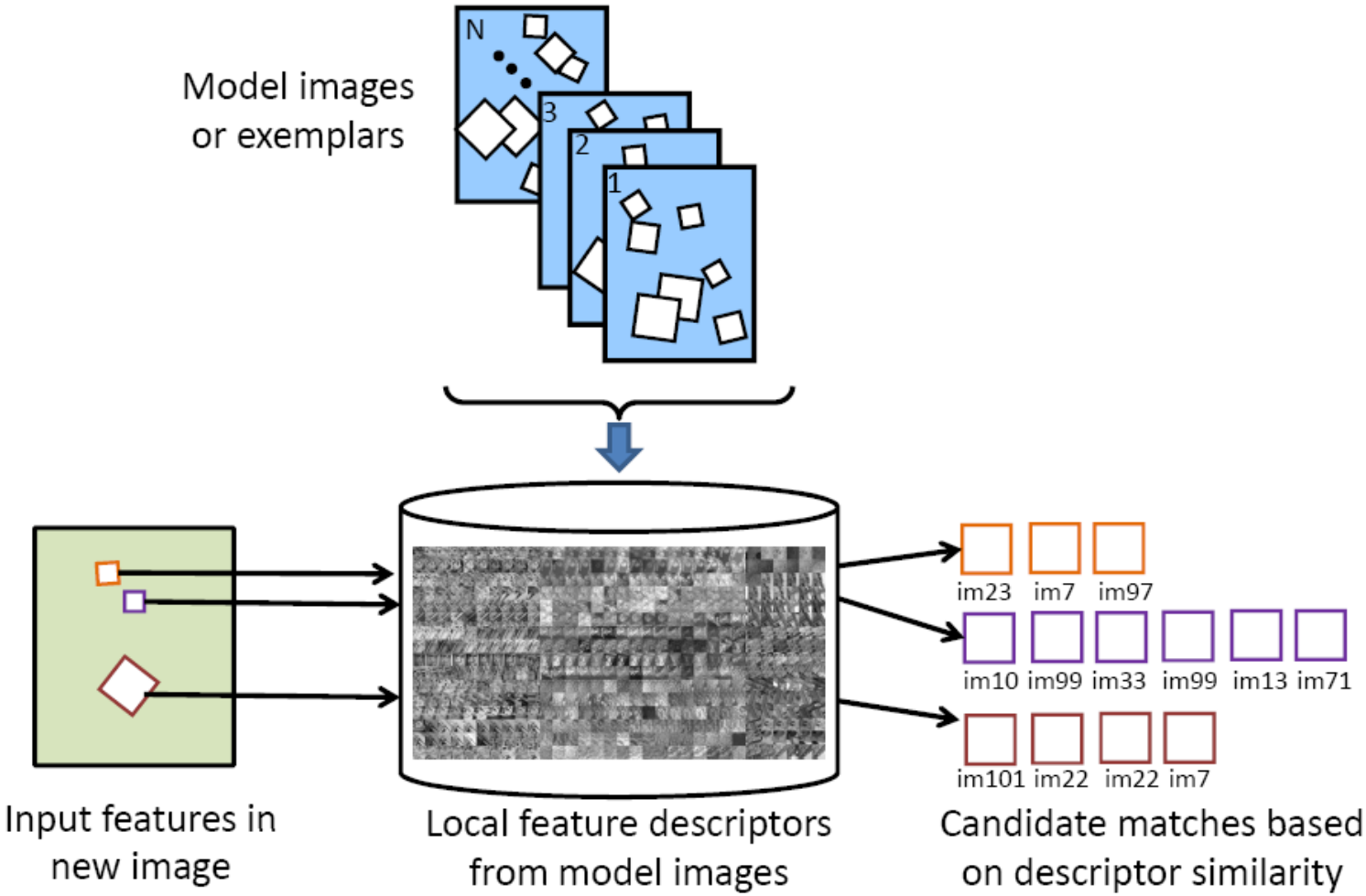


Image credit: K. Grauman and B. Leibe

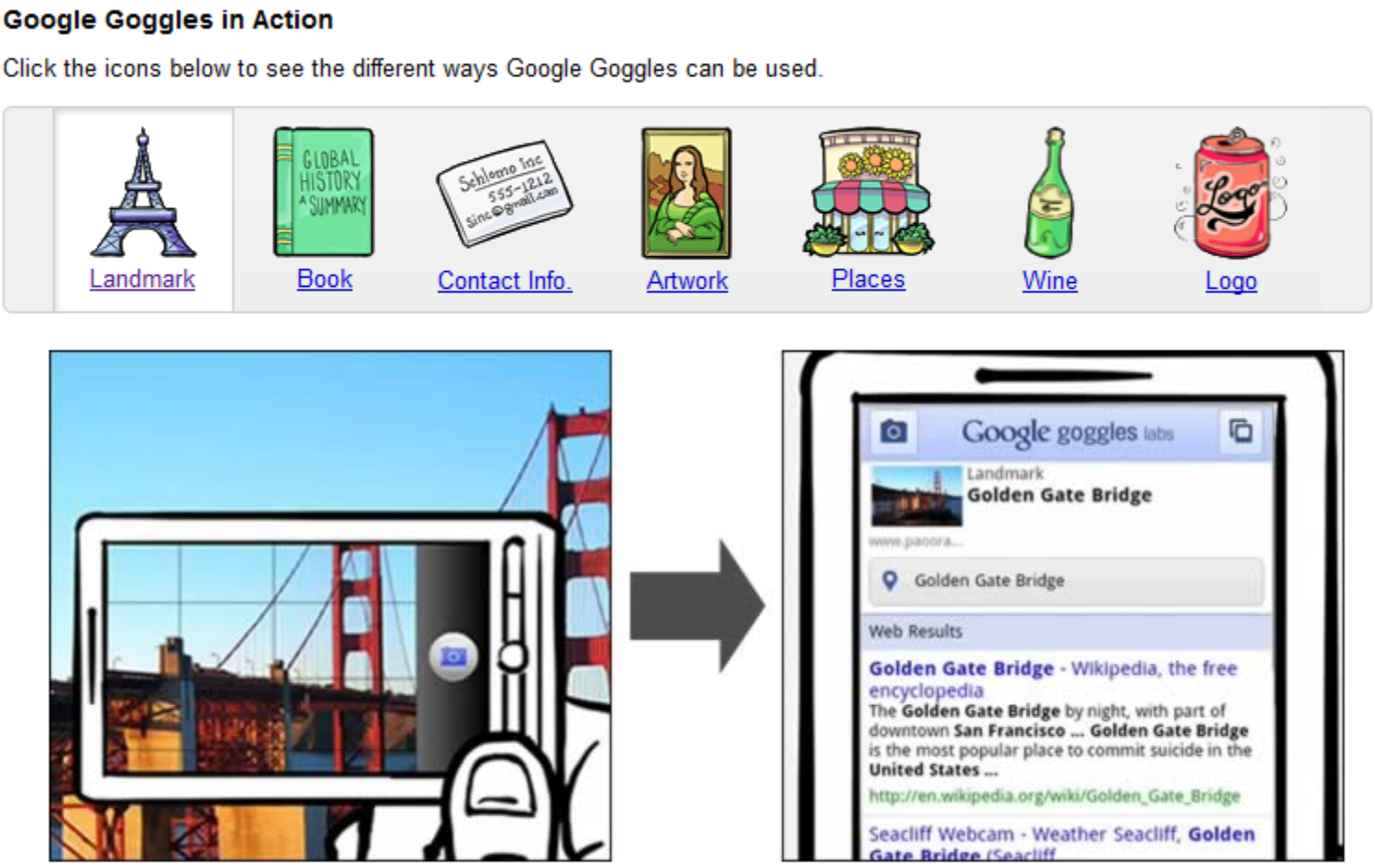
Large-scale image search

Combining local features, indexing, and spatial constraints

Philbin et al. ‘07

Large-scale image search

Combining local features, indexing, and spatial constraints

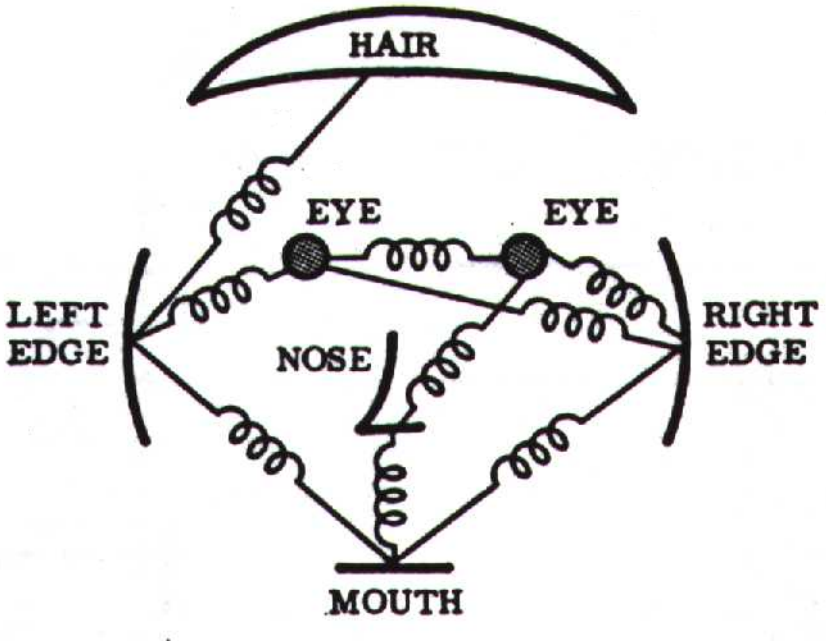


History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- 1990s – present: sliding window approaches
- Late 1990s: local features
- **Early 2000s: parts-and-shape models**

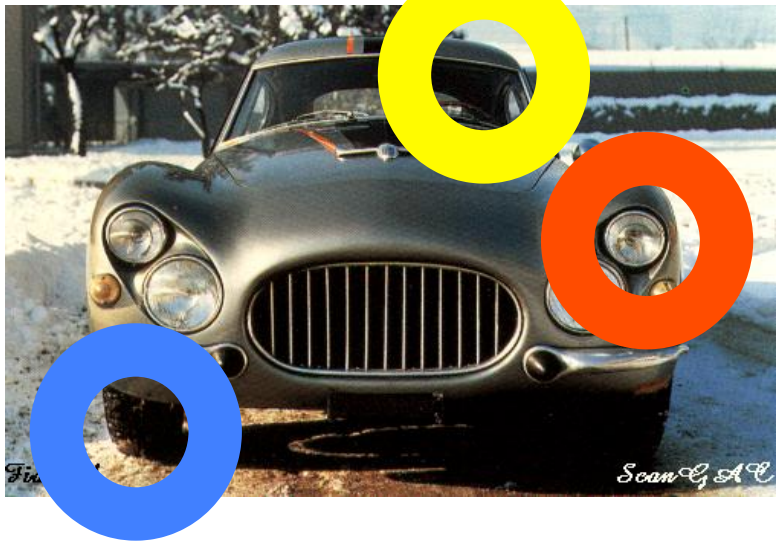
Parts-and-shape models

- Model:
  - Object as a set of parts
  - Relative locations between parts
  - Appearance of part



Fischler & Elschlager 73

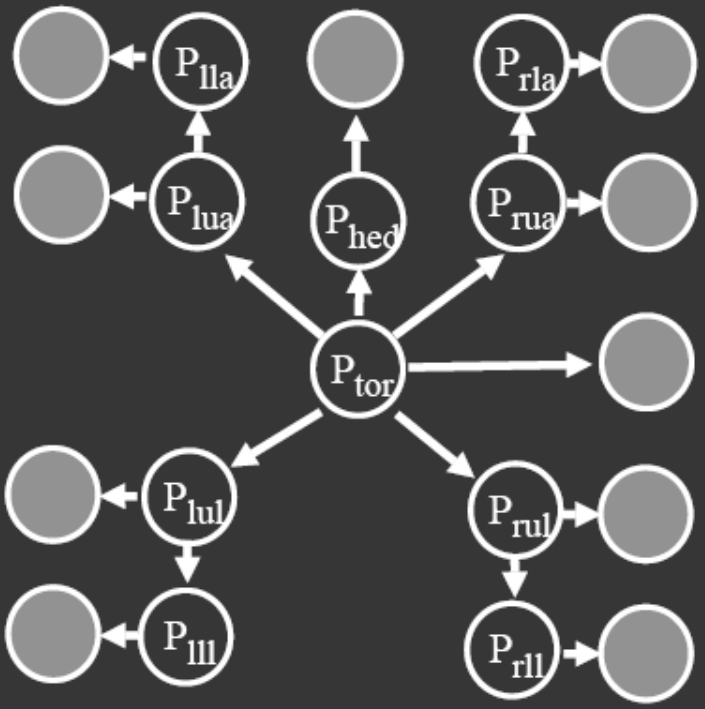
Constellation models

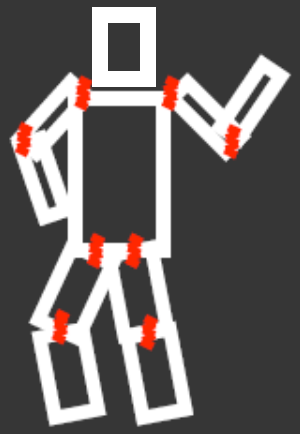


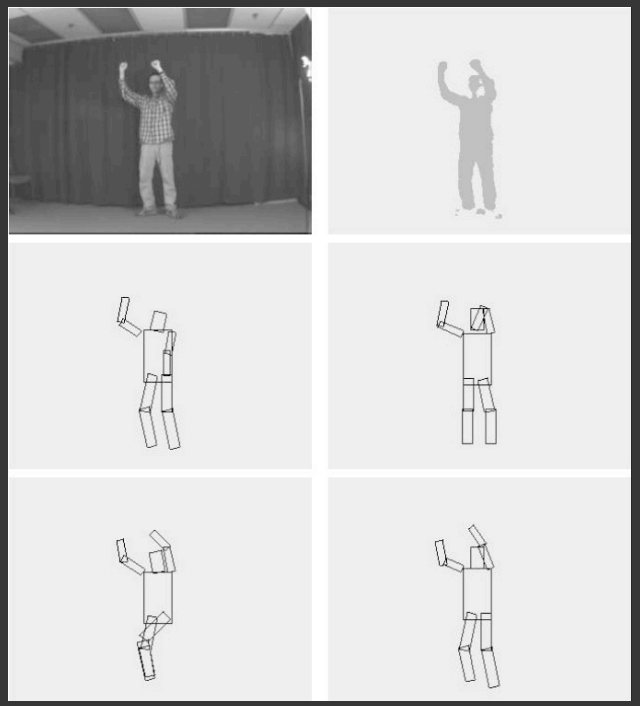
Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

Pictorial structure model

Fischler and Elschlager(73), Felzenszwalb and Huttenlocher(00)





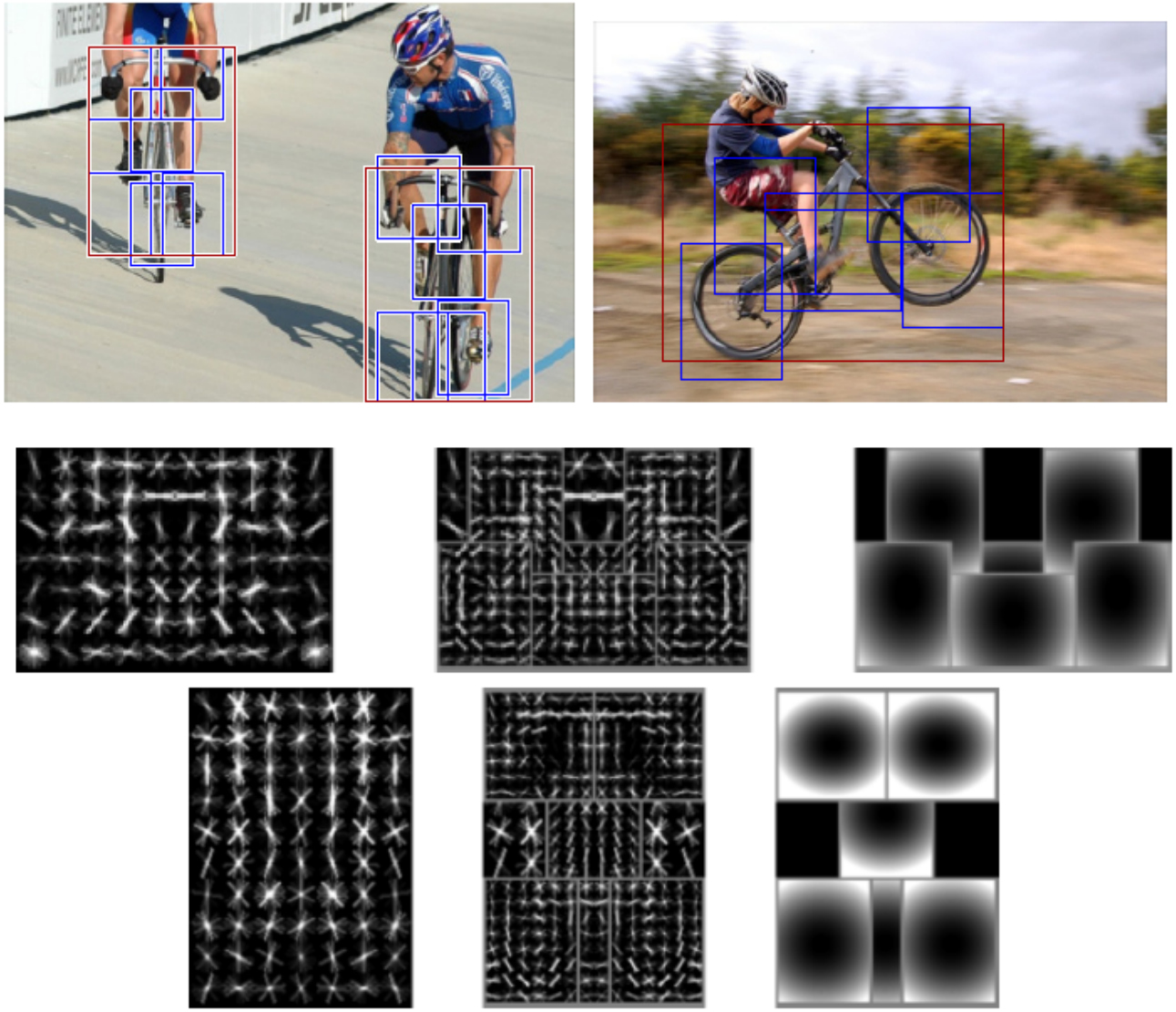


$$\Pr(P_{tor}, P_{arm}, \ldots | \mathrm{Im}) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(\mathrm{Im}(P_i))$$

part geometry

part appearance

Discriminatively trained part-based models



P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan,
"Object Detection with Discriminatively Trained Part-Based Models," PAMI 2009

History of ideas in recognition

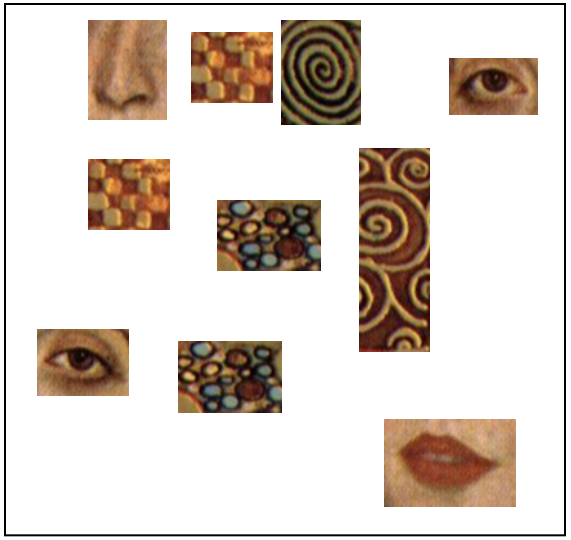
- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- 1990s – present: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- **Mid-2000s: bags of features**

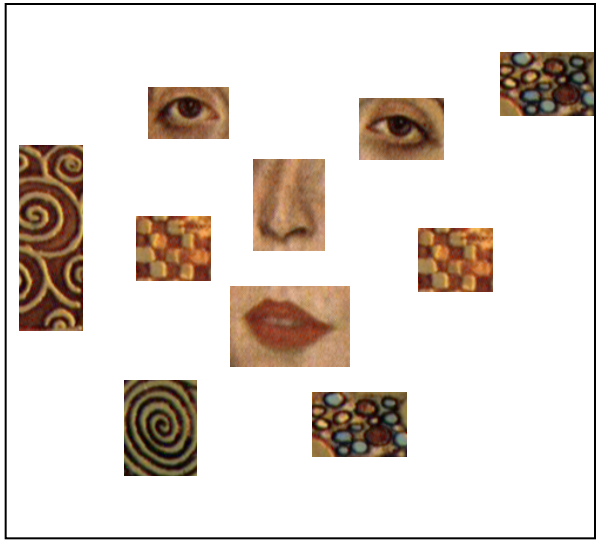
Bag-of-features models

Object → **Bag of ‘words’**

Objects as texture

- All of these are treated as being the same





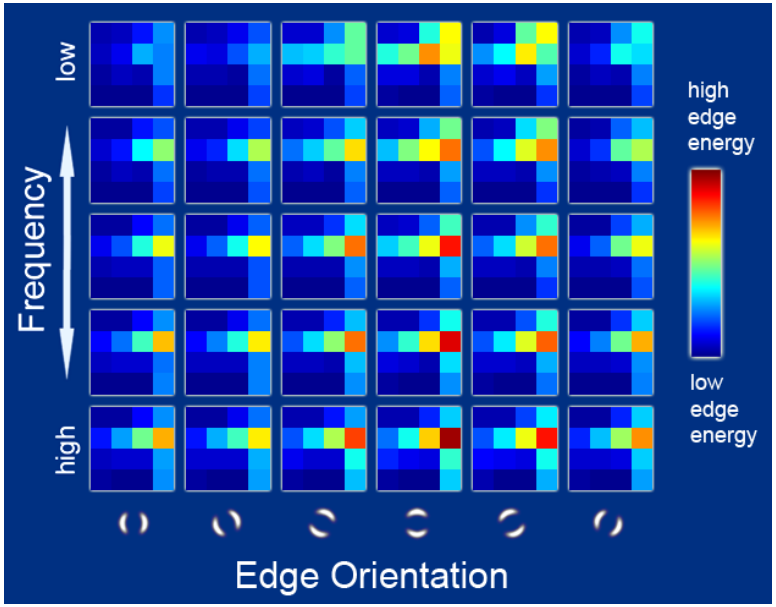
- No distinction between foreground and background: scene recognition?

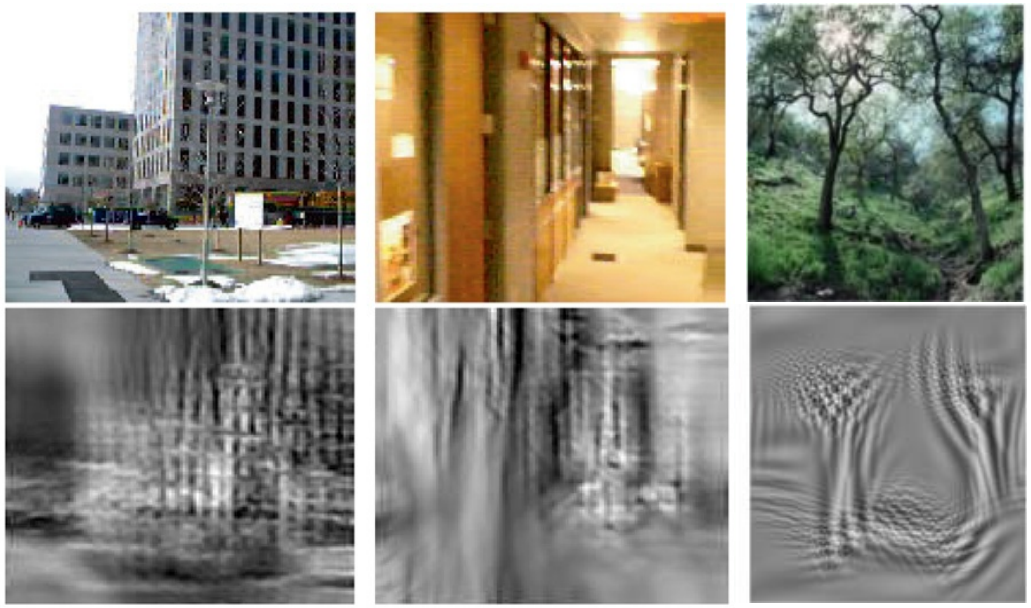
History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- 1990s – present: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features
- **Present trends: “big data”, context, attributes, combining geometry and recognition, advanced scene understanding tasks, deep learning and convolutional networks**

Global appearance models revisited

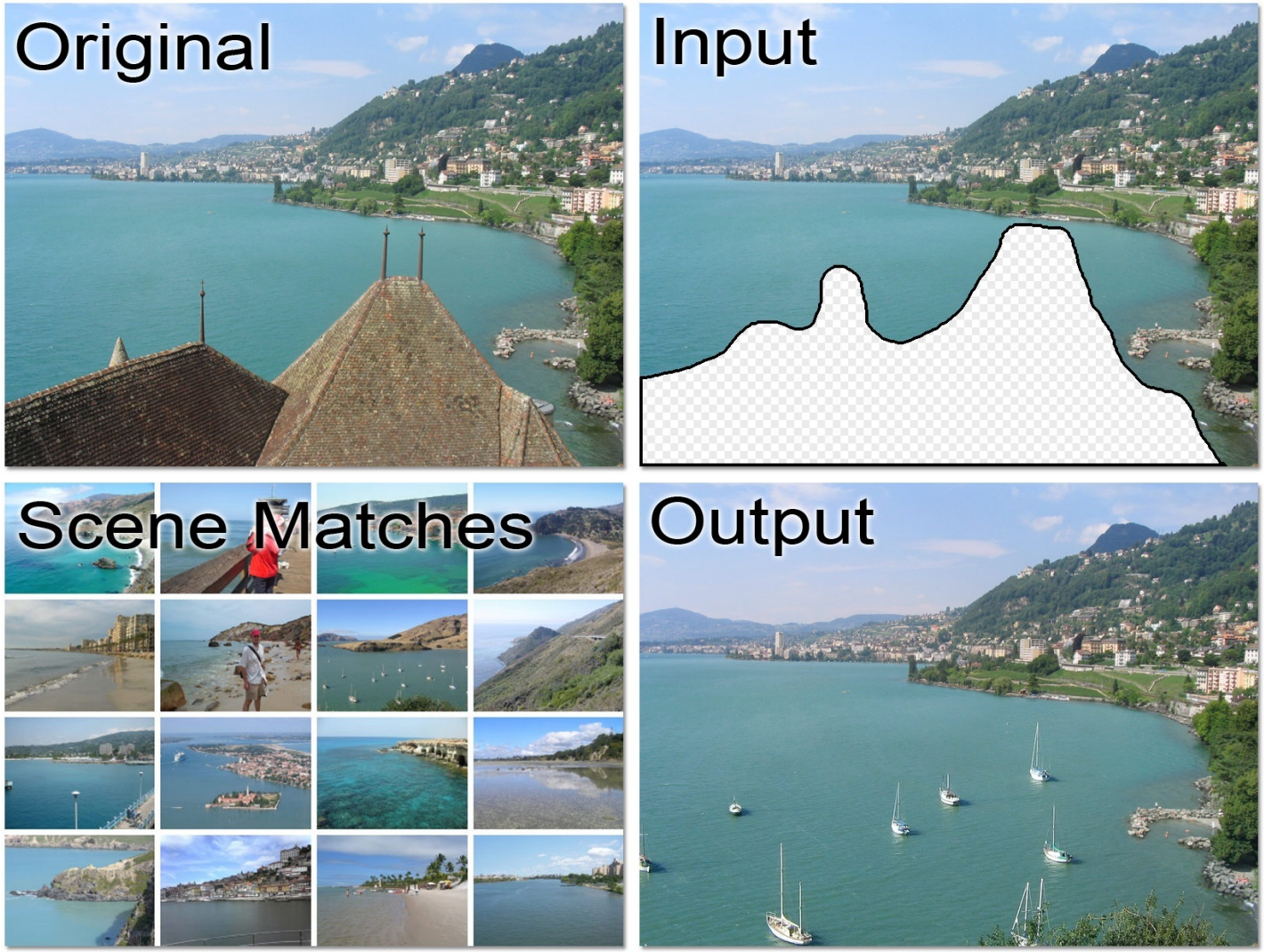
- The “gist” of a scene: Oliva & Torralba (2001)





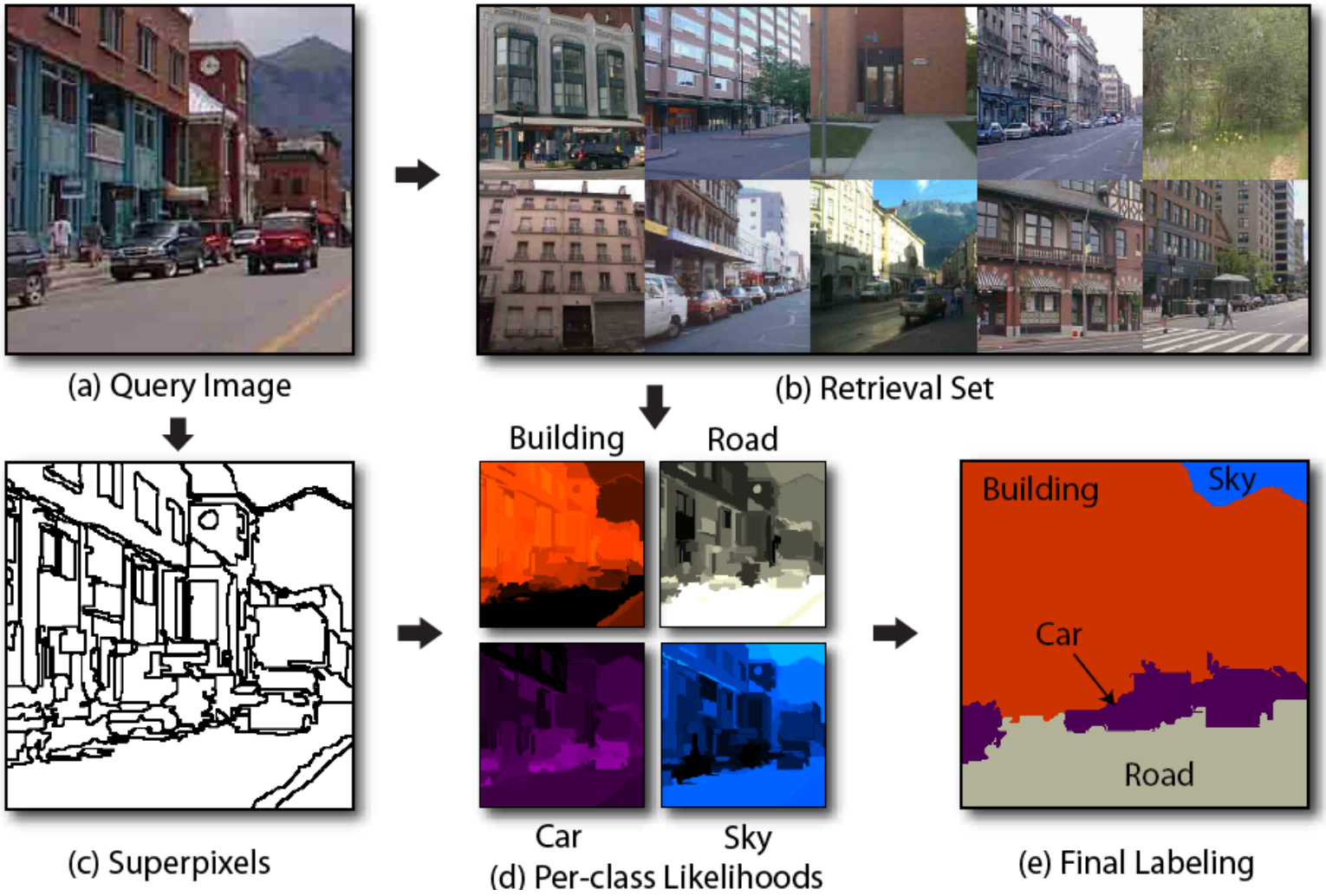
http://people.csail.mit.edu/torralba/code/spatialenvelope/

Data-driven methods



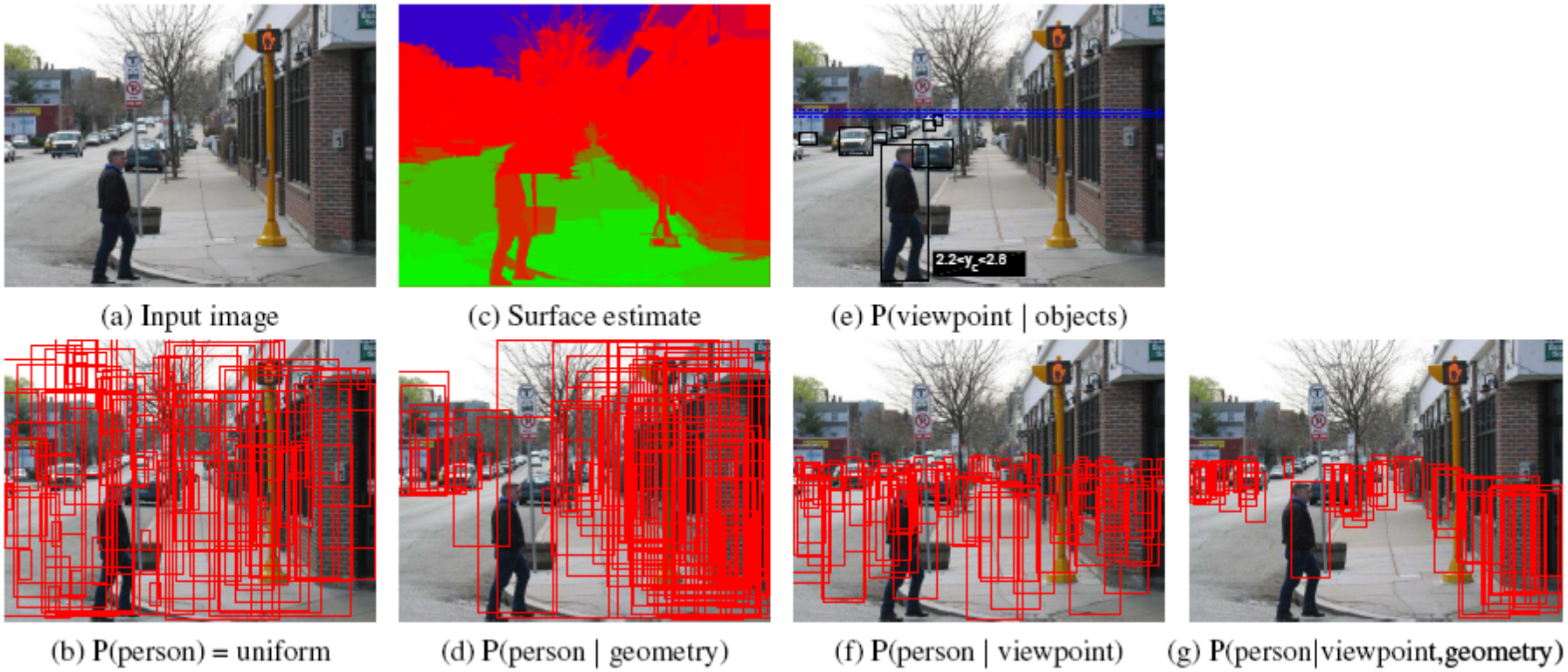
J. Hays and A. Efros, Scene Completion using Millions of Photographs, SIGGRAPH 2007

Data-driven methods



J. Tighe and S. Lazebnik, ECCV 2010

Geometric context



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

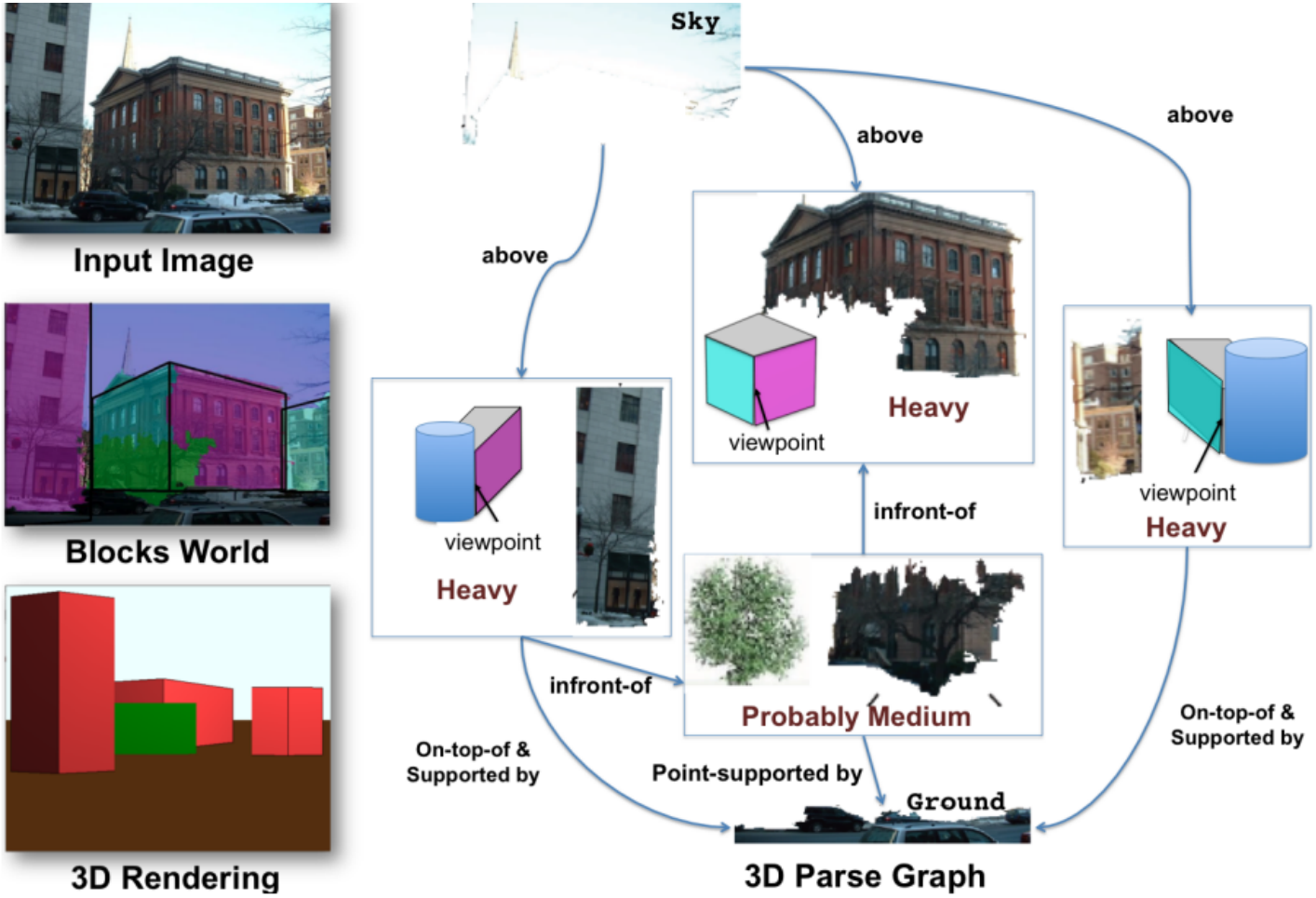
(b) P(person) = uniform (d) P(person | geometry) (f) P(person | viewpoint) (g) P(person|viewpoint,geometry)

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

Geometry and recognition

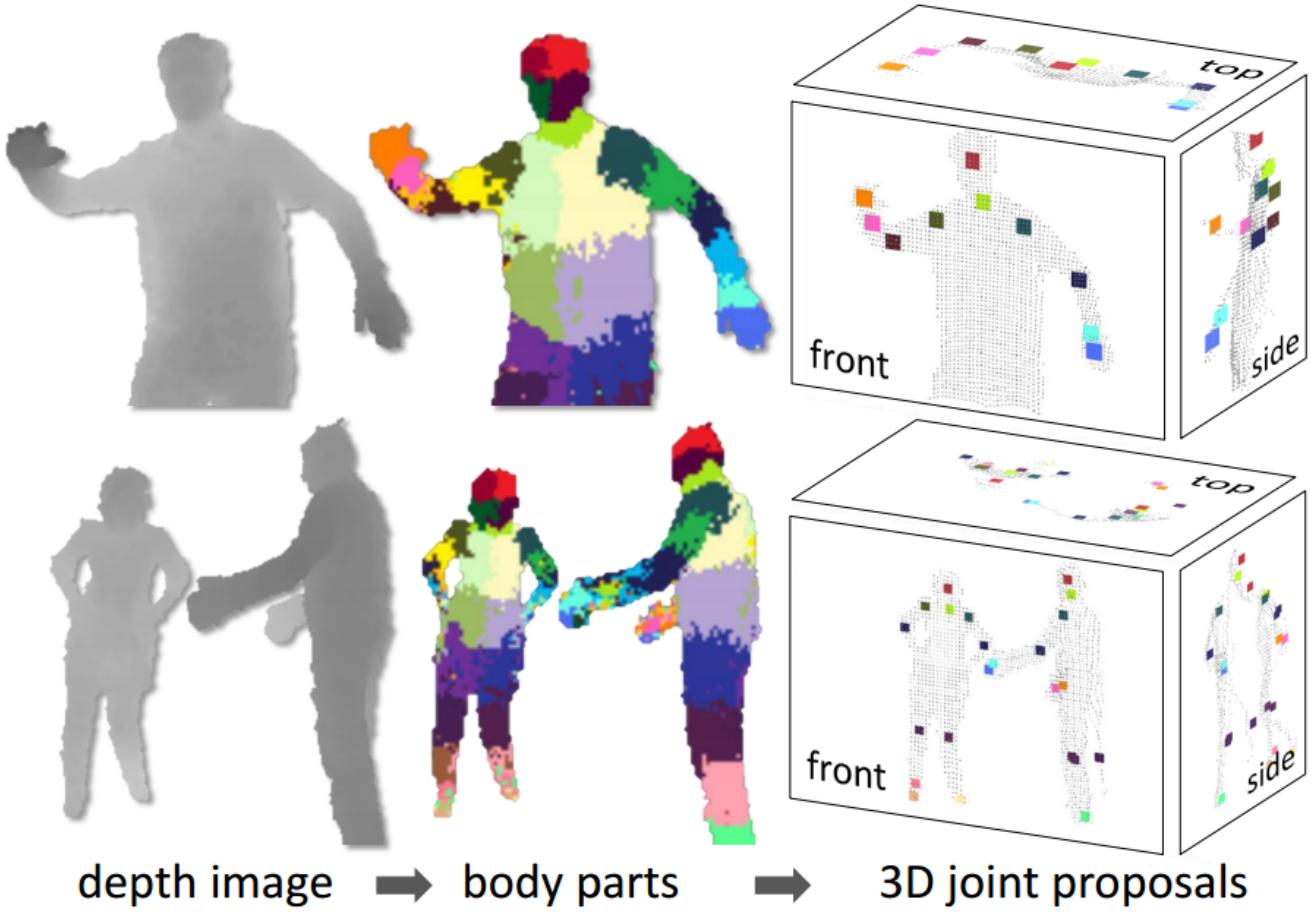
V. Hedau, D. Hoiem, and D. Forsyth,
Recovering the Spatial Layout of Cluttered Rooms, ICCV 2009.

Geometry and recognition



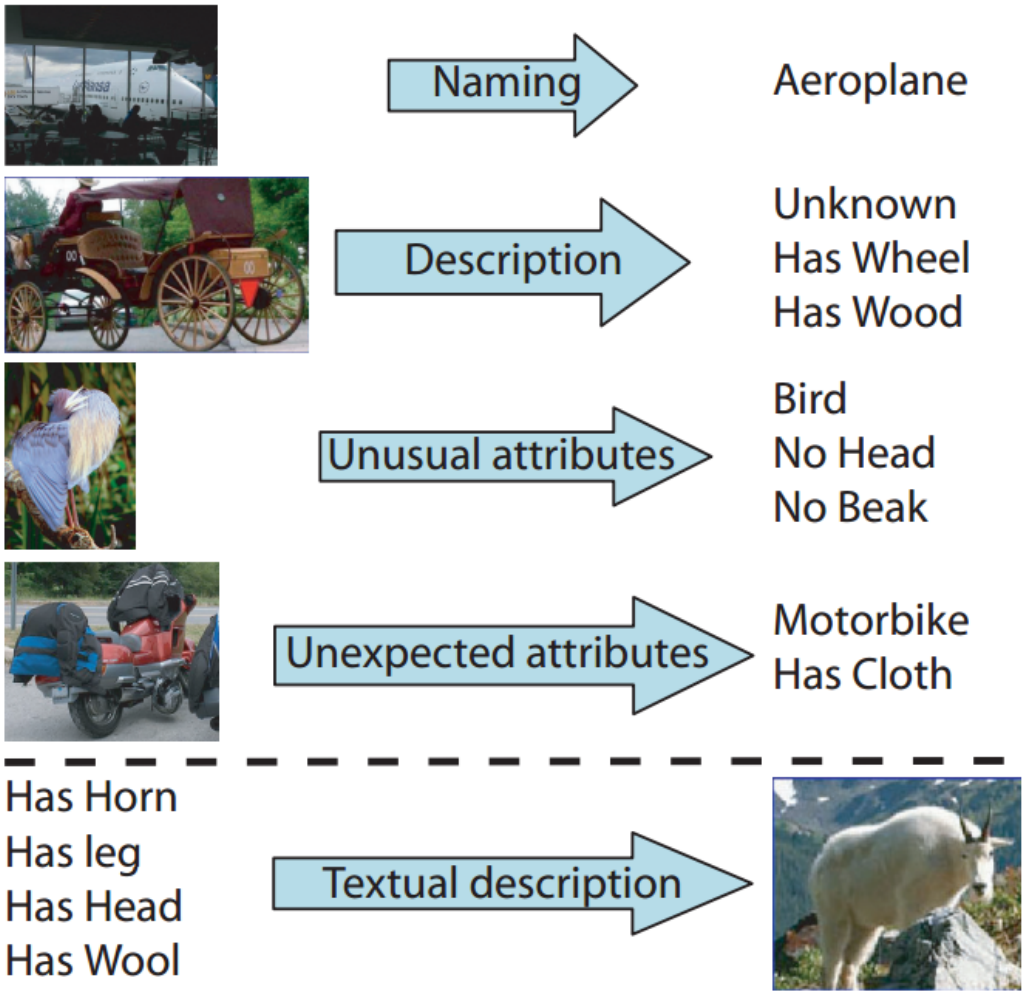
A. Gupta, A. Efros and M. Hebert,
[Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics](), ECCV 2010

Recognition from RGBD Images



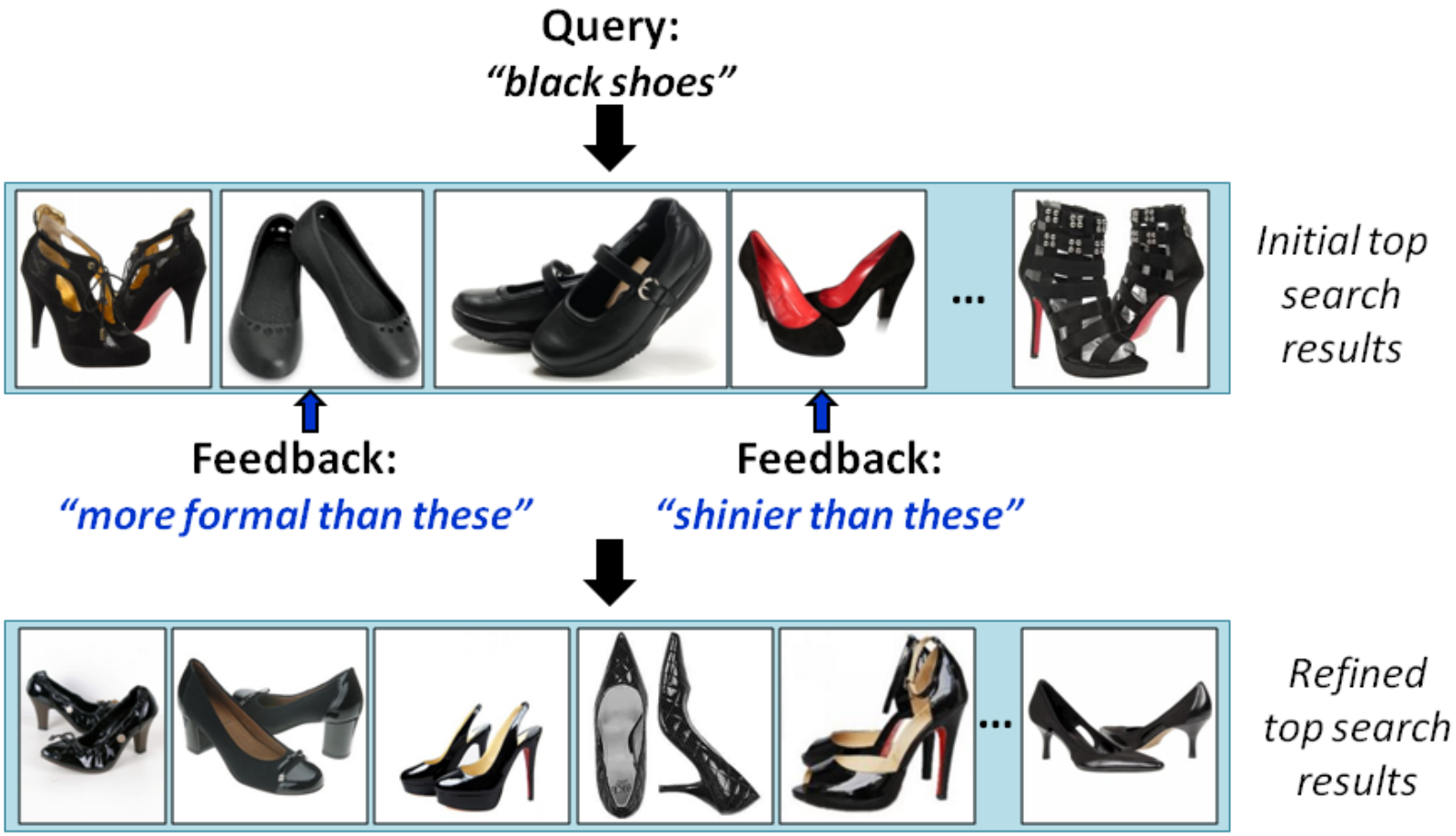
J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, **Real-Time Human Pose Recognition in Parts from a Single Depth Image**, CVPR 2011

Attributes for recognition



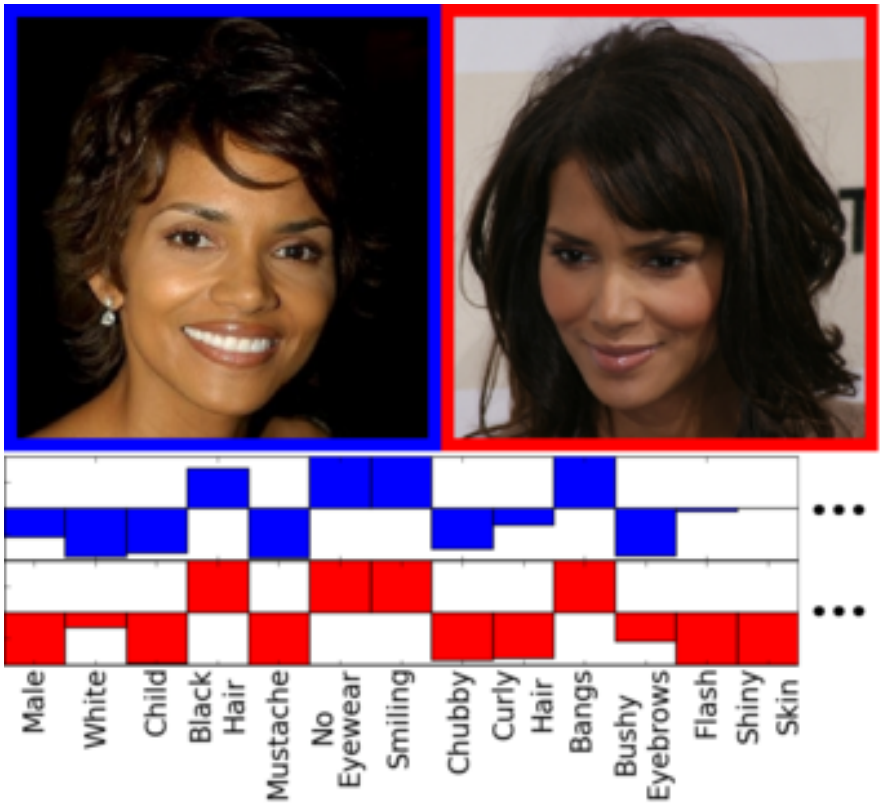
A. Farhadi, I. Endres, D. Hoiem, and D Forsyth,
Describing Objects by their Attributes, CVPR 2009

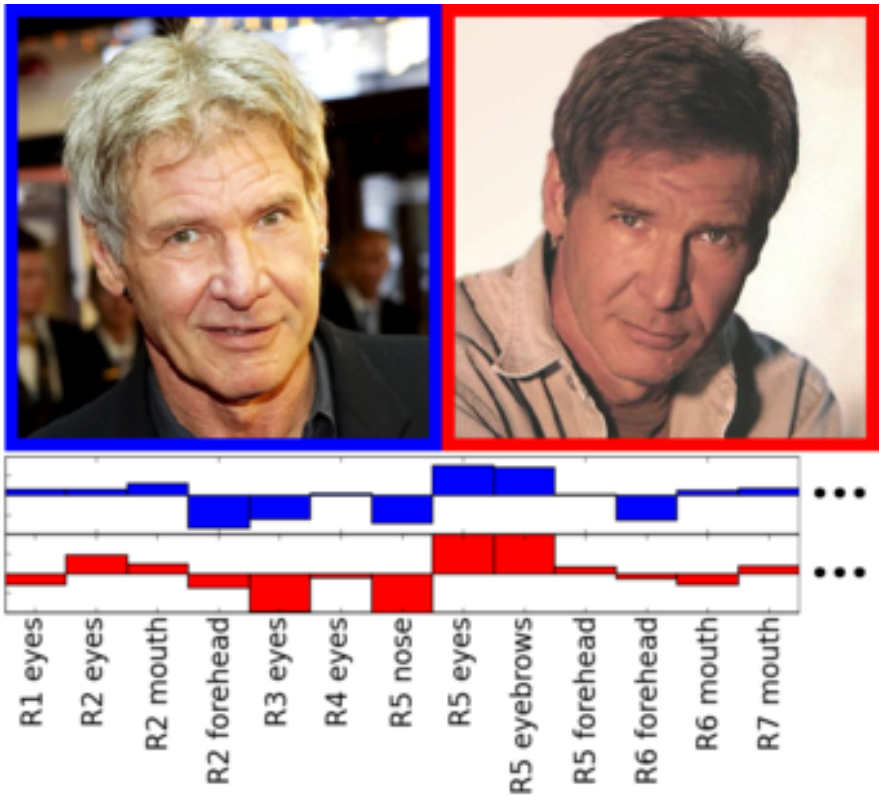
Attributes for visual search



A. Kovashka, D. Parikh and K. Grauman,
WhittleSearch: Image Search with Relative Attribute Feedback, CVPR 2012

Attributes for face verification





N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar,
Attribute and Simile Classifiers for Face Verification, ICCV 2009

Advanced image understanding: Sentence generation from images

This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.

This is a picture of two dogs. The first dog is near the second furry dog.

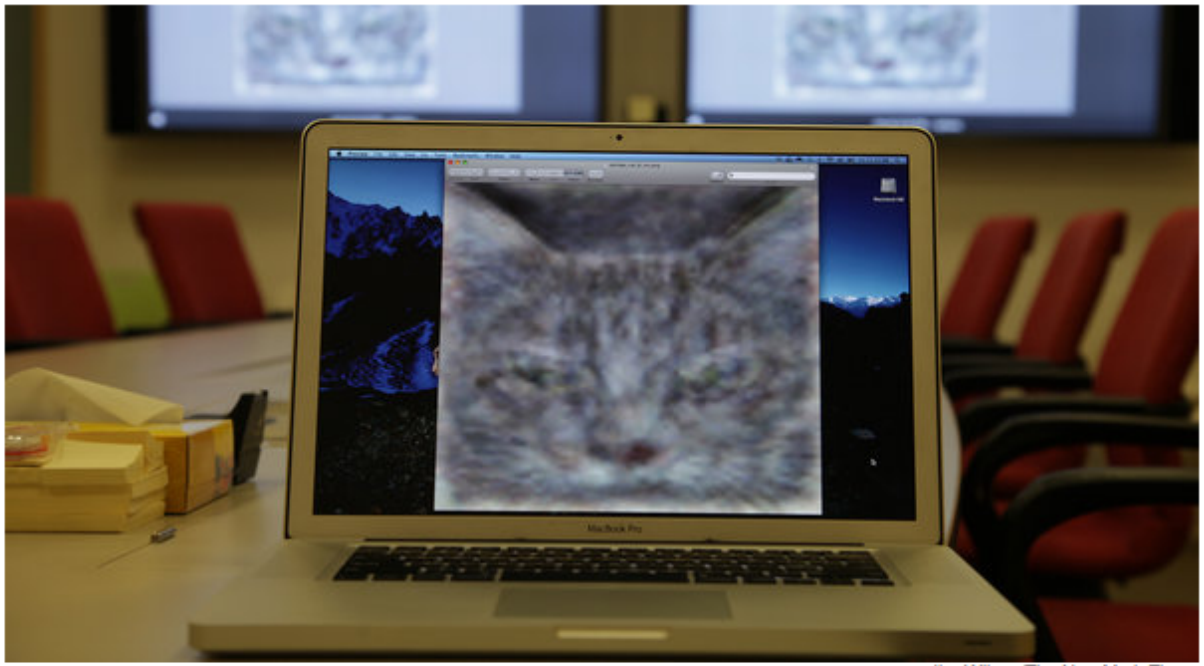
G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, T. Berg,
Baby Talk: Understanding and Generating Simple Image Descriptions, CVPR 2011

Deep learning

The New York Times

Business Day
Technology

How Many Computers to Identify a Cat? 16,000



Jim Wilson/The New York Times

An image of a cat that a neural network taught itself to recognize.

By JOHN MARKOFF
Published: June 25, 2012

[NY Times article]