

SAIR

Spatial AI & Robotics Lab

CSE 473/573-A

L18: RECOGNITION

Chen Wang

Spatial AI & Robotics Lab

Department of Computer Science and Engineering



University at Buffalo The State University of New York



Objects vs Texture vs Scene

The texture



The object



The scene



Scenes vs. objects

A photograph of a firehydrant



A photograph of a street



Scene views vs. objects

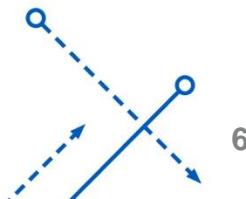
- Scene:
 - a place in which a human can act within
 - or a place to which a human being could navigate.



- A lot more than just a combination of objects
 - just as objects are more than the combinations of their parts.
- Associated with specific functions and behaviors (like objects)
 - E.g., eating in a restaurant, reading in a library,
 - Talking in classroom, playing in a park, etc

Why be concerned about the difference?

- Features to be extracted?
- Context for Recognition?
- Various **categories** have features that can be essential for interpretation.



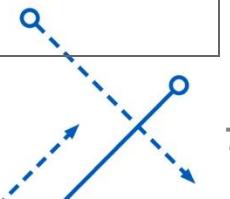
Why do we care about categories?

Perception of function:

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

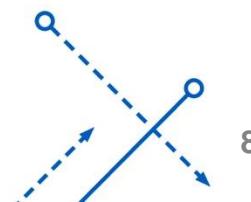


“We therefore include the perception of function as a proper –indeed, crucial- subject for vision science”, from *Vision Science, chapter 9, Palmer*.



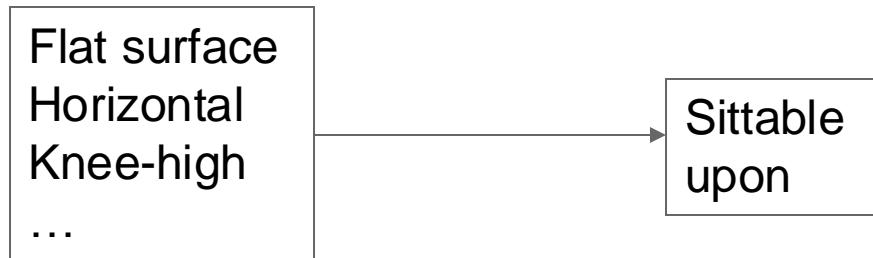
Why do we care about categories?

- When we recognize an object, we can make
 - predictions about its behavior in the future
 - beyond of what is immediately perceived.



The perception of function

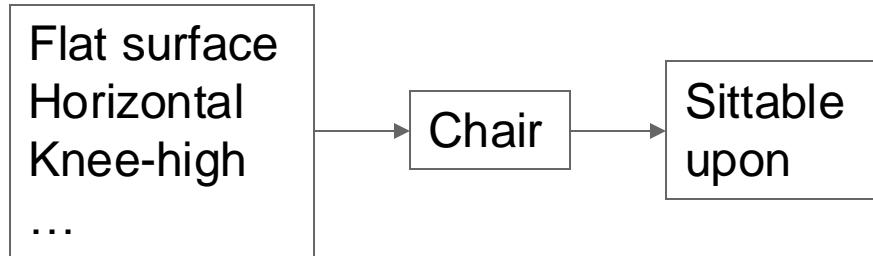
- Direct perception (affordances): [introduced by Gibson]



Chair Chair



- Mediated perception (Categorization)



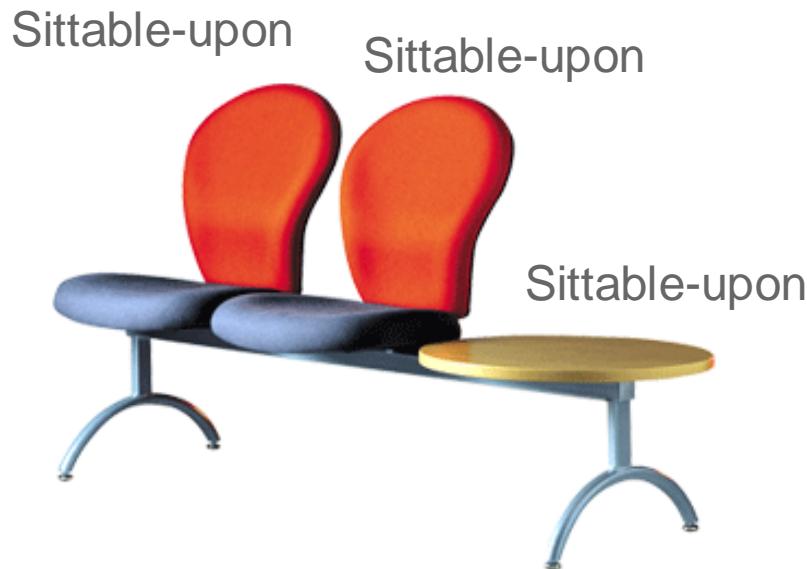
- One caveat of this comparison:
 - Deciding that something is a chair might require access to more features than deciding that we can sit on something
 - A different level of categorization.

Direct perception

Some aspects of an object function can be perceived directly

- Functional form:

- Some forms clearly indicate to a function
- e.g., “sittable-upon”, container, cutting device, ...



It does not seem easy
to sit-upon this...



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...

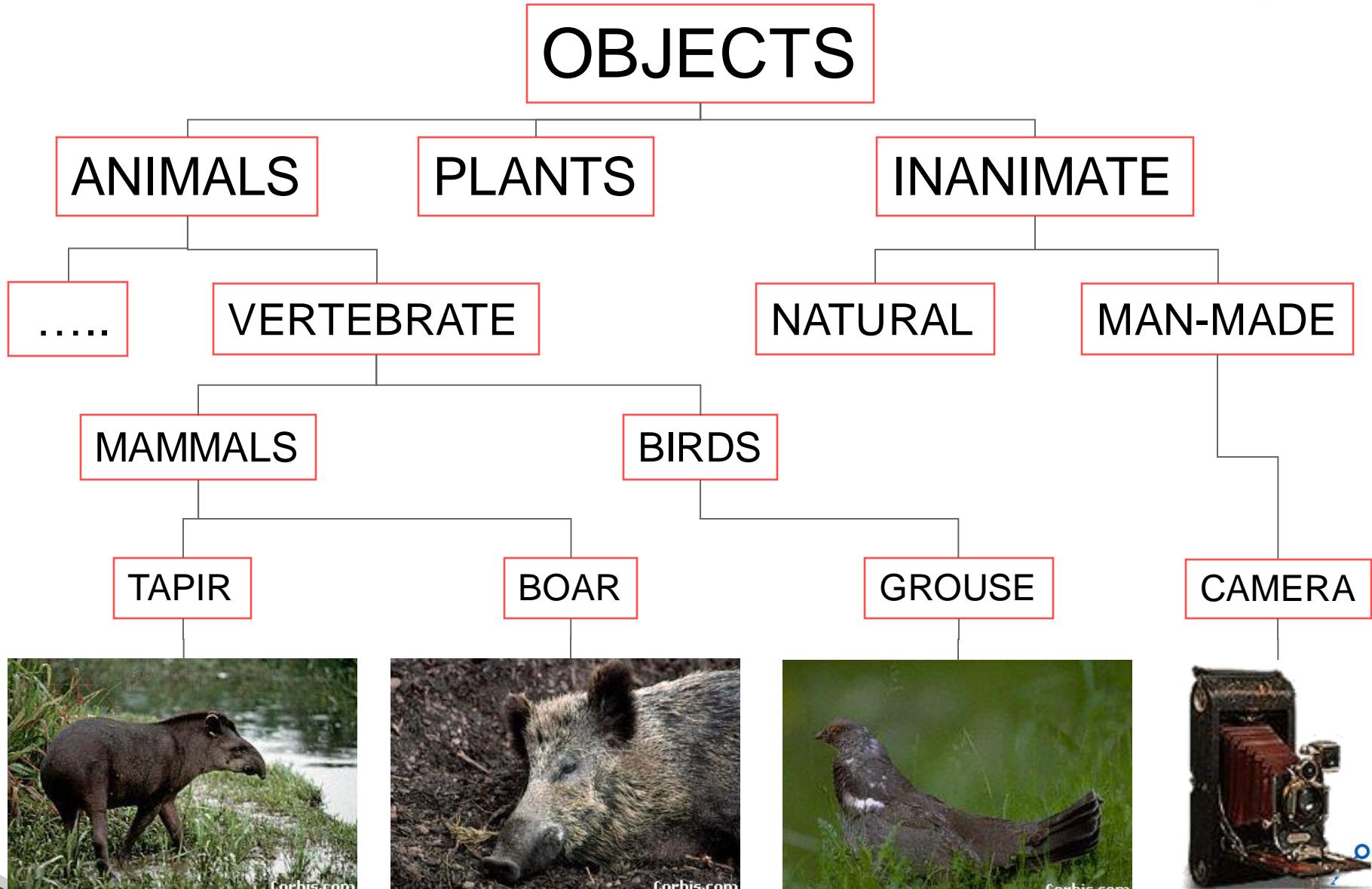
Which level of categorization is the right one?

- Car is an object composed of:
 - a few doors, four wheels (not always visible), a roof
 - front lights, windshield



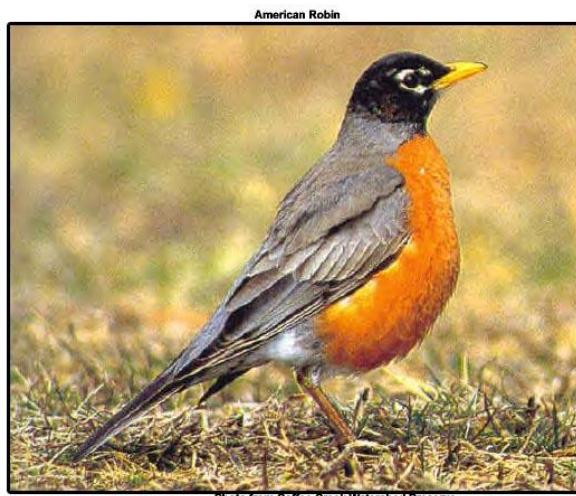
If you are thinking in buying a car, you might want to be a bit more specific about your categorization.

Object Category



Entry-level categories

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



A bird



An ostrich

(Jolicoeur, Gluck, Kosslyn 1984)

Demo : Rapid image understanding

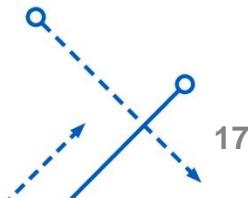
Instructions: 9 photographs will be shown for half a second each. Your task is to **memorize these pictures**.

GET READY

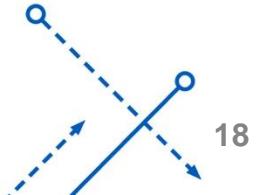
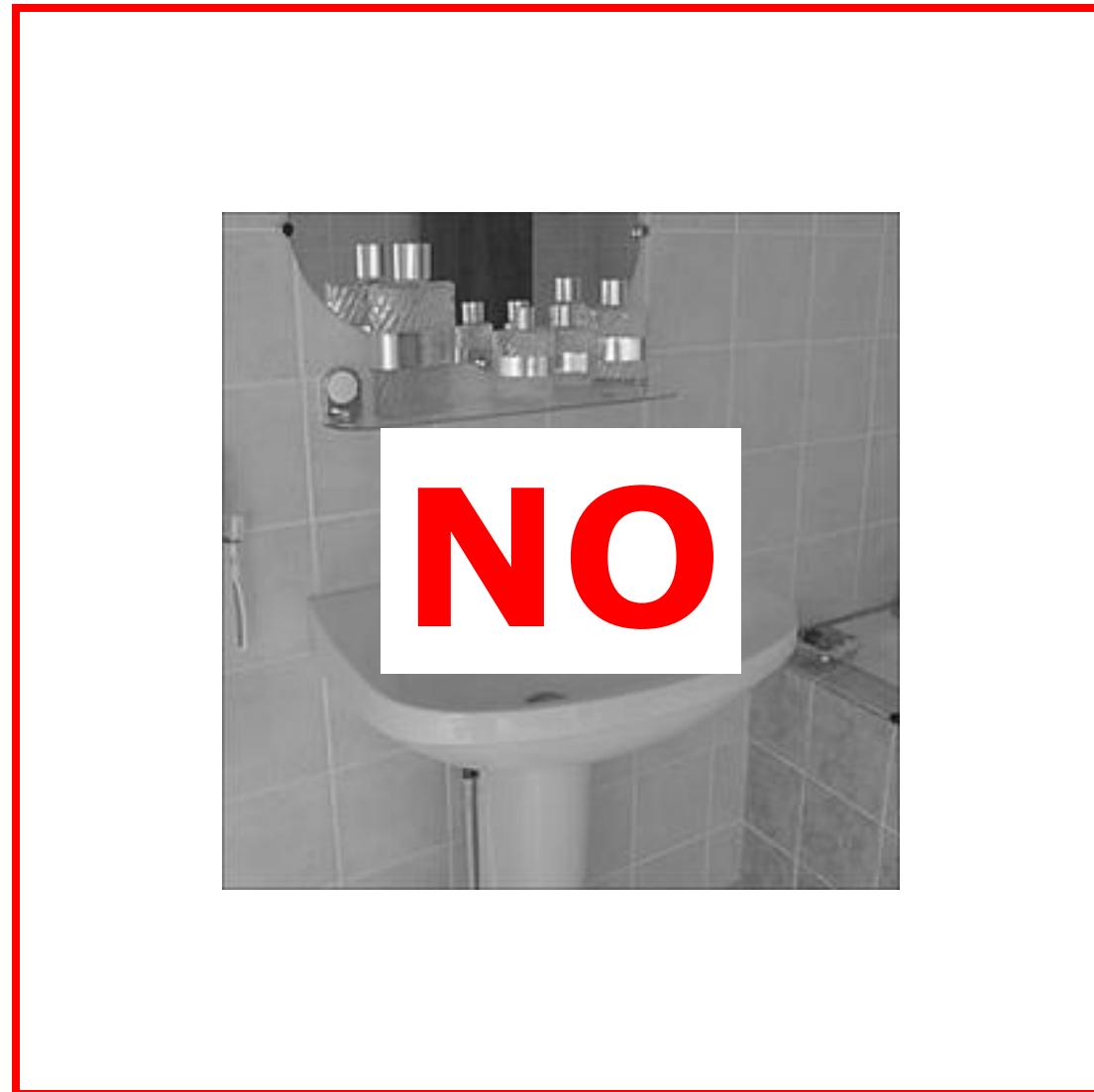
Memory Test



Have you seen this picture ?



Memory Test

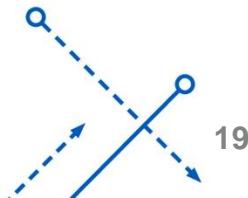


18

Memory Test



Have you seen this picture ?



Memory Test



Memory Test



Have you seen this picture ?

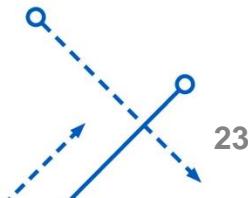
Memory Test



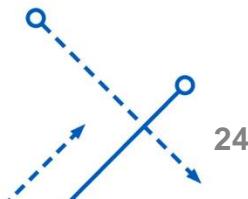
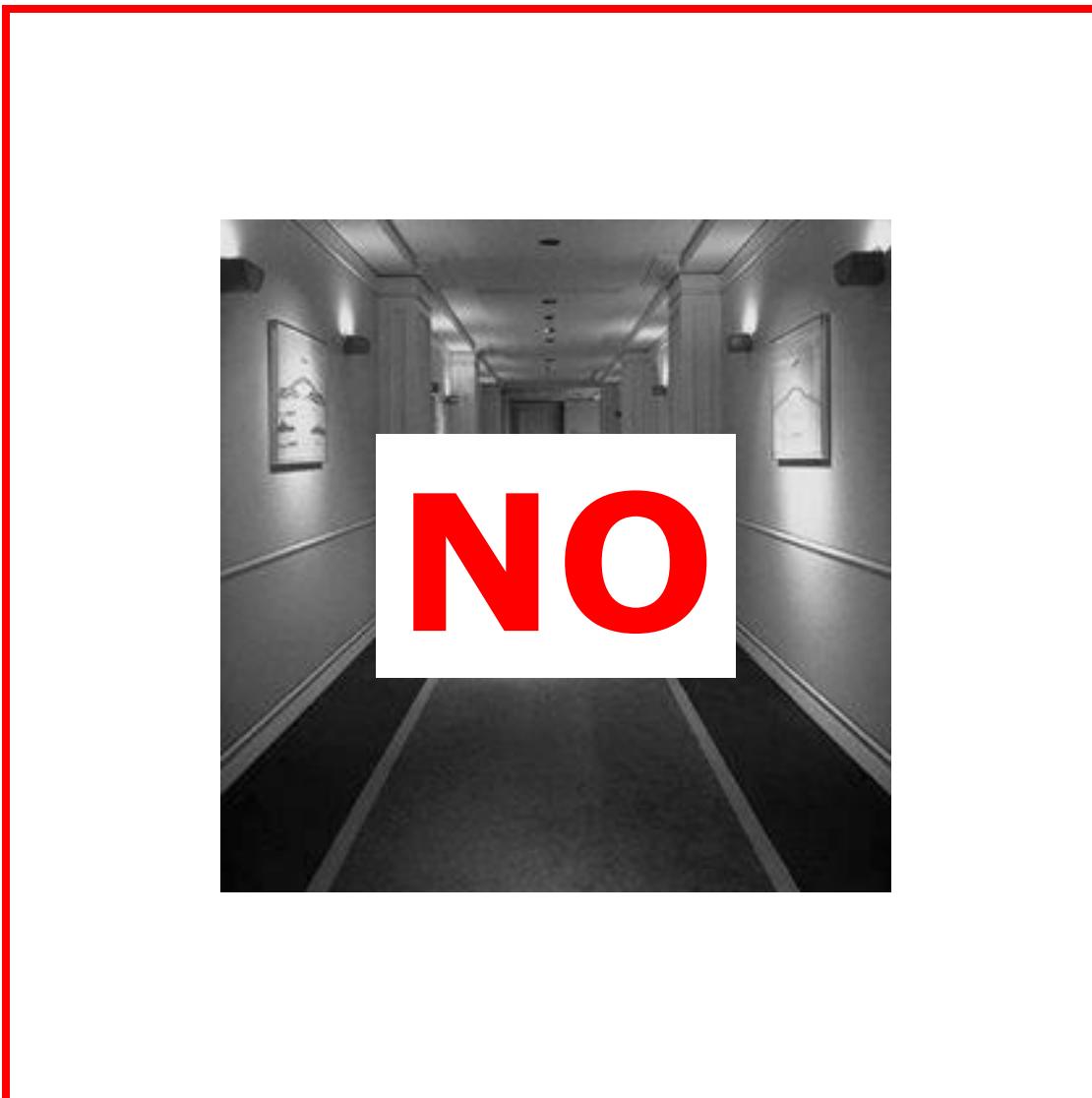
Memory Test



Have you seen this picture ?



Memory Test



Memory Test



Have you seen this picture ?

Memory Test



Memory Test



Have you seen this picture ?

Memory Test

o



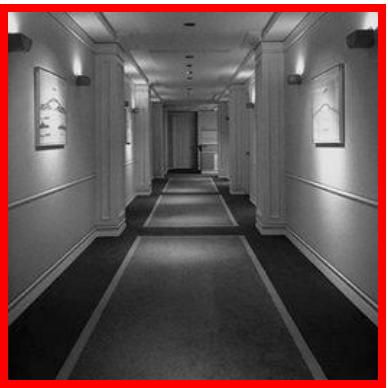
o
o
28

Memory Test Overview

You have seen these pictures



You were tested with these pictures



The gist of the scene

In a glance, we remember the meaning of an image and its global **layout** but some objects and **details** are forgotten



Mary Potter (1976)

Mary Potter (1975, 1976) demonstrated

- during a rapid sequential visual presentation (100 msec per image), a novel picture is instantly **understood** and observers seem to comprehend a lot of visual information



Object recognition: Is it really so hard?

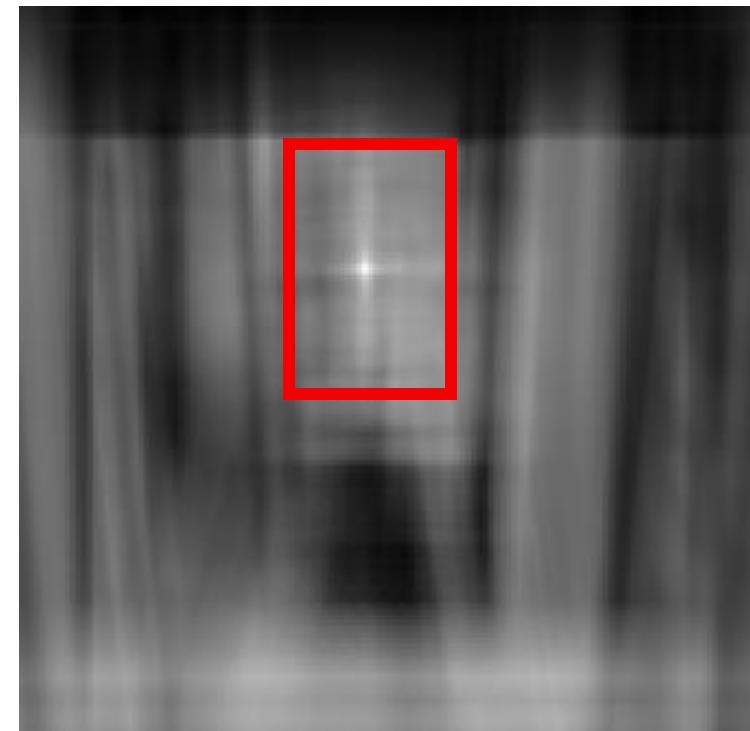
This is a chair



Find the chair in this image

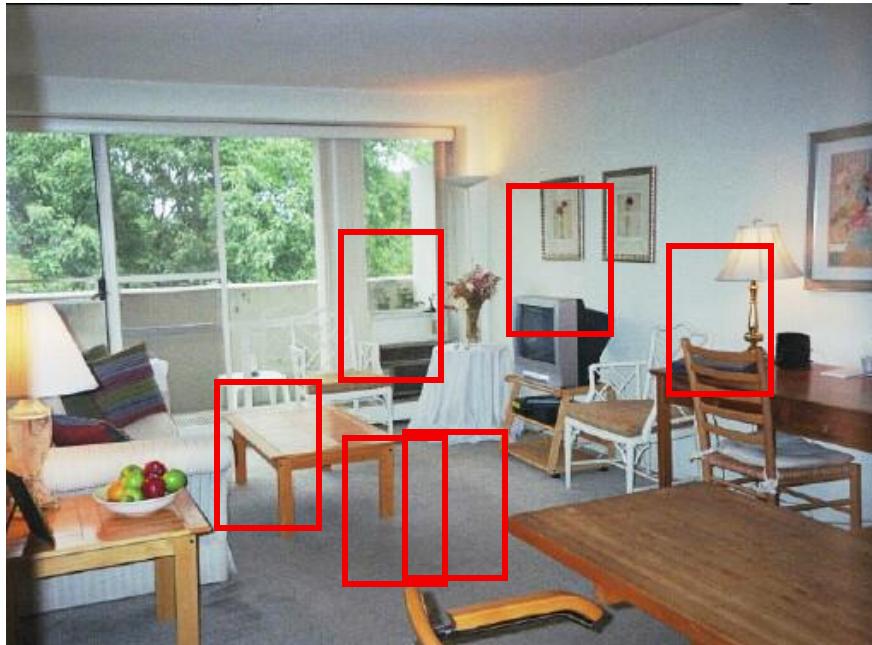


Output of normalized correlation



Object recognition: Is it really so hard?

Find the chair in this image



Pretty much garbage
Template matching cannot make it.

Challenges: robustness



Illumination



Object pose



Clutter



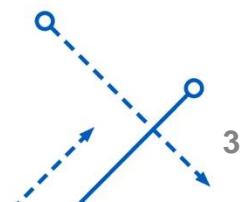
Occlusions



Intra-class
appearance

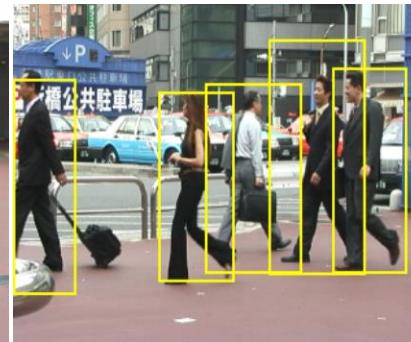
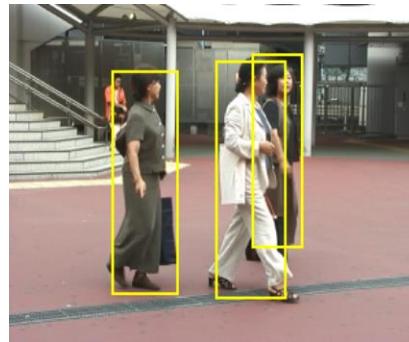
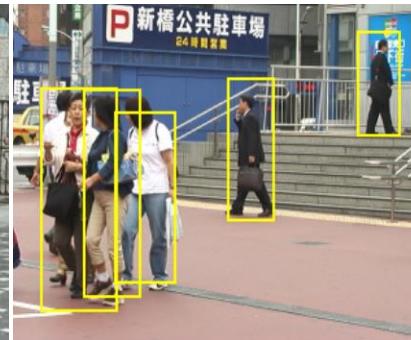
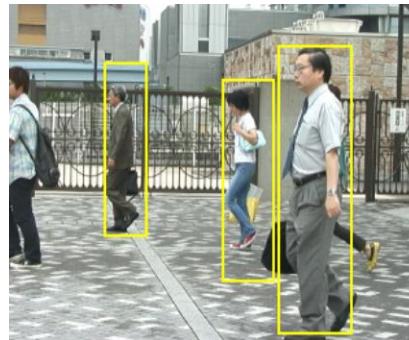


Viewpoint

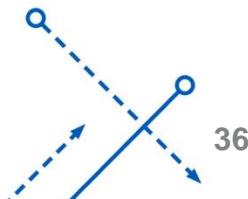


Challenges: robustness/invariance

- Detection in Crowded Scenes
 - Learn object variability
 - Changes in appearance, scale, and articulation
 - Compensate for clutter, overlap, and occlusion



Challenges: context and human experience

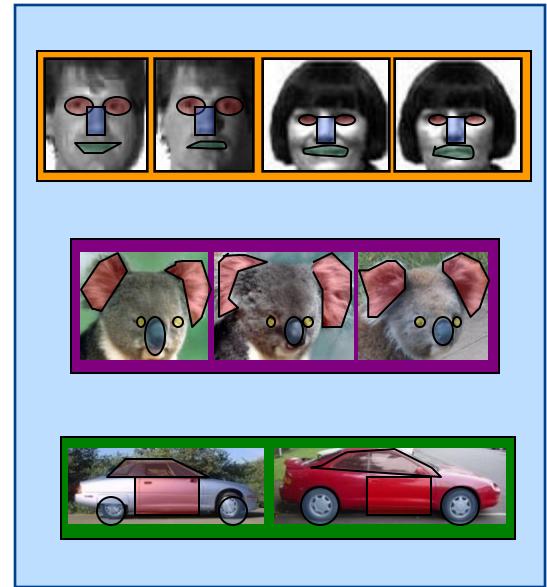
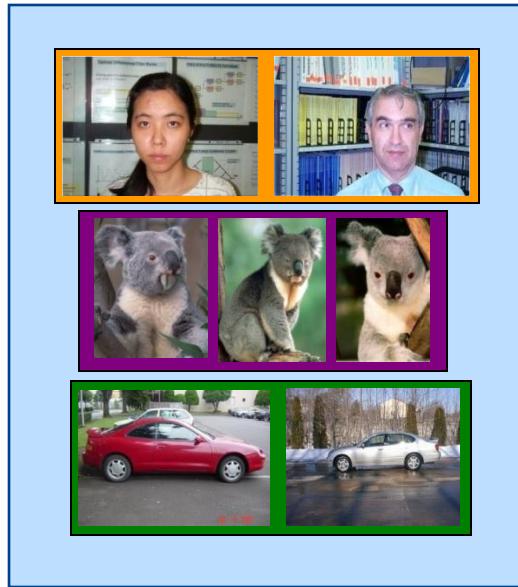


Challenges: learning with minimal supervision

Less



More



Unlabeled,
multiple objects

Classes labeled,
some clutter

Cropped to object,
parts and classes
labeled

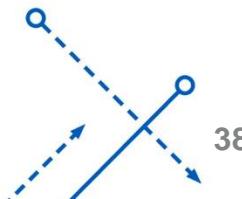
Inputs/outputs/assumptions

What is the **goal**?

- Say yes/no as to whether an object present in image

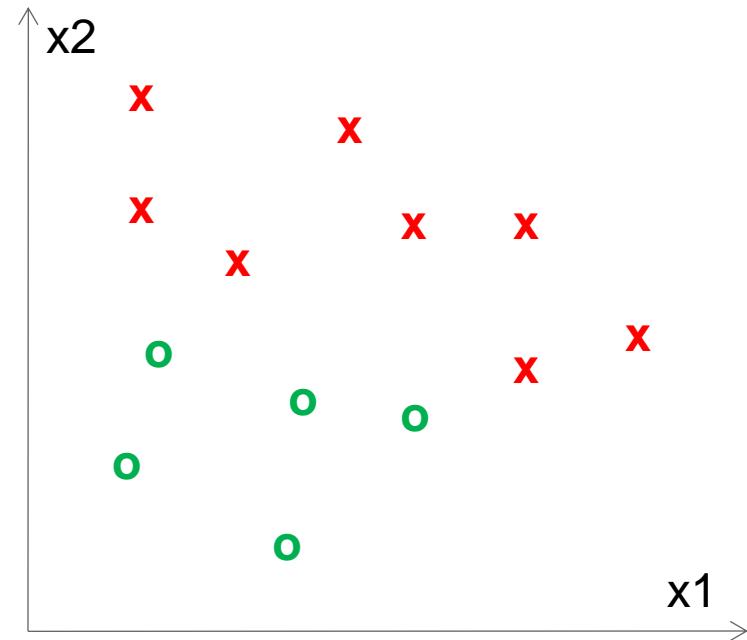
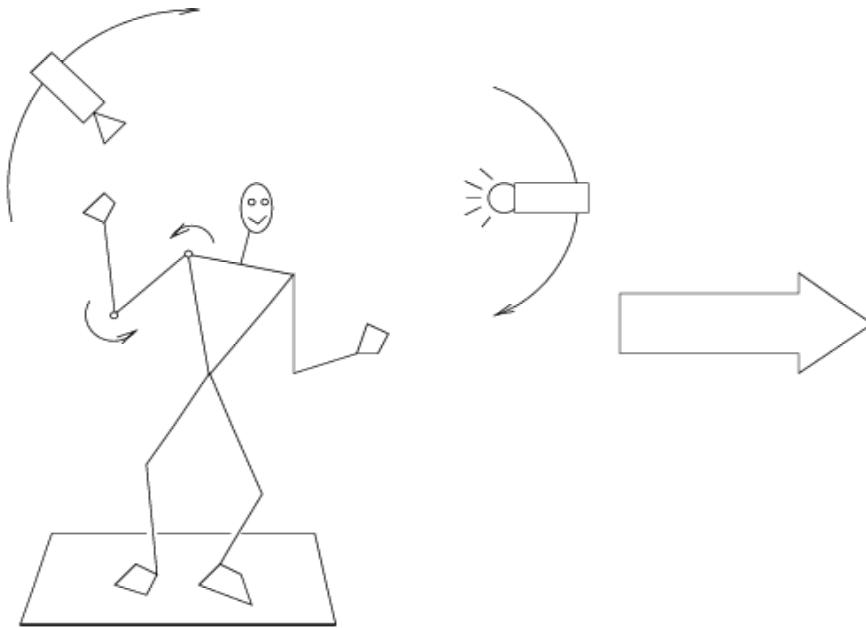
And/or:

- Categorize all objects
- Forced choice from pool of categories
- Bounding box on object
- Full segmentation
- Build a model of an object category
- Determine pose of an object, e.g., for robot to grasp



Recognition is all about modeling variability

- Invariance is the key challenge to computer vision.



Variability:
Camera position
Illumination
Shape parameters

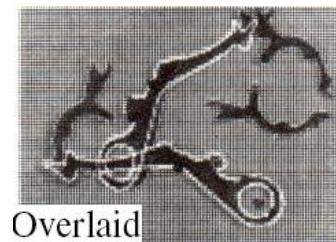
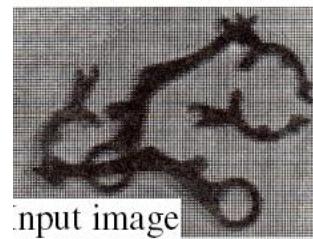
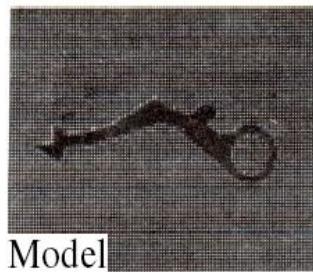
Within-class variations?



Within-class variations



Rough Evolution of Focus in Object Recognition



1980s



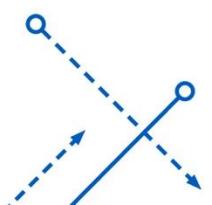
7 5 9 2 6 5
1 2 2 2 2 3
0 2 3 8 0 7



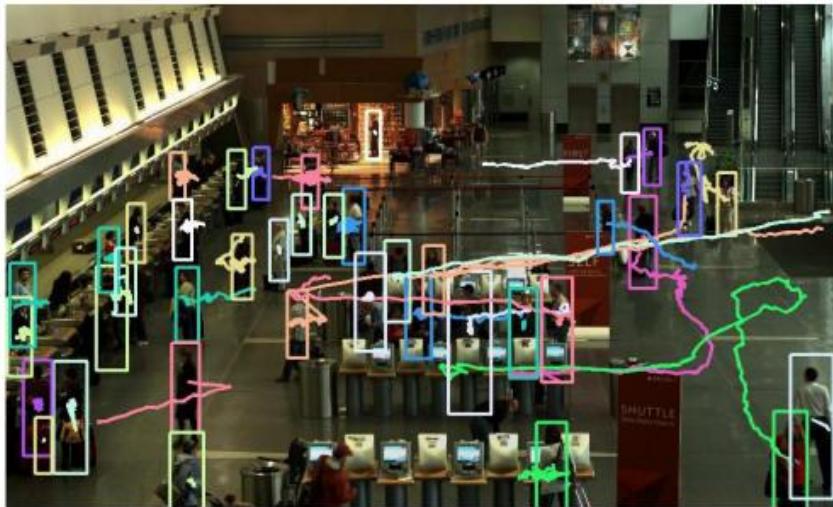
1990s to early 2000s



2000-2010...



Today



In 2014, image recognition by computer outperforms human.

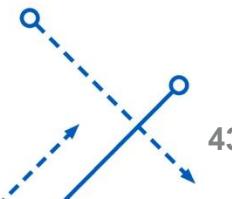
History of ideas in recognition

- 1960s – early 1990s: the geometric era

Variability: Camera position
Illumination Focus: **Alignment**

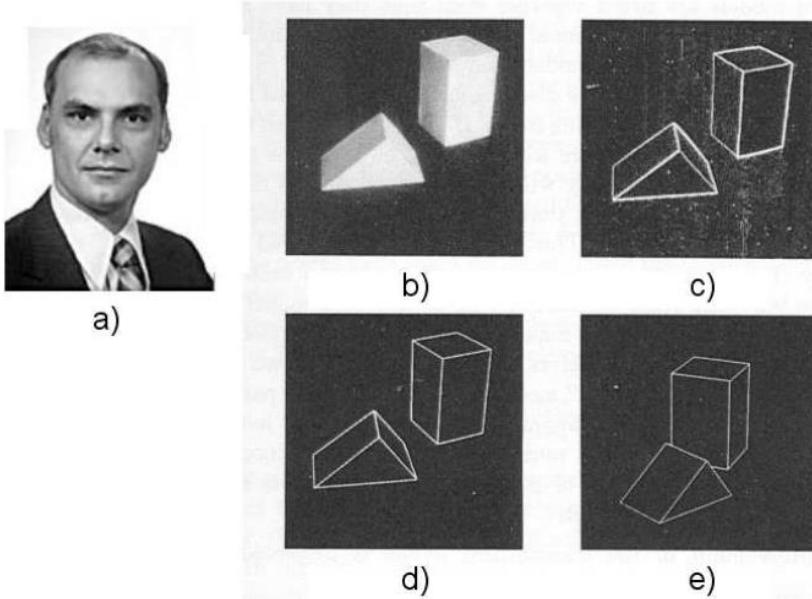
Shape: Assumed known

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



History of ideas in recognition

- Recognition as an alignment problem: Block world



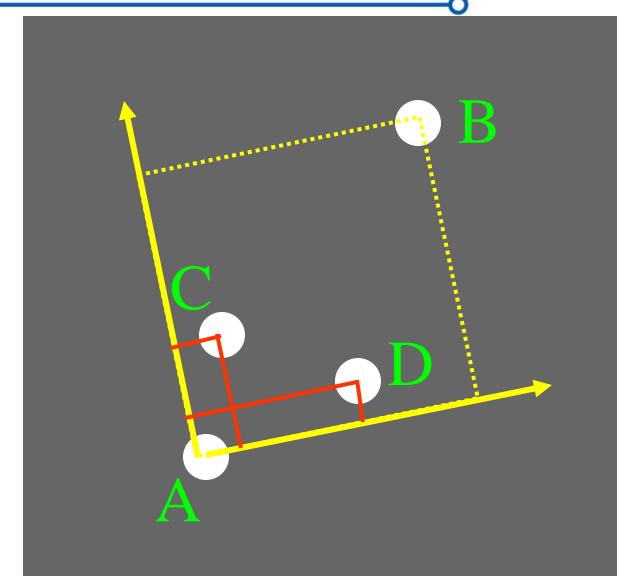
L. G. Roberts, [Machine Perception of Three Dimensional Solids](#), Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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.)

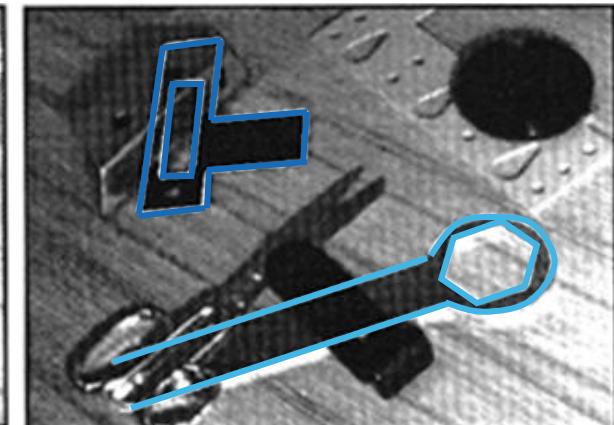
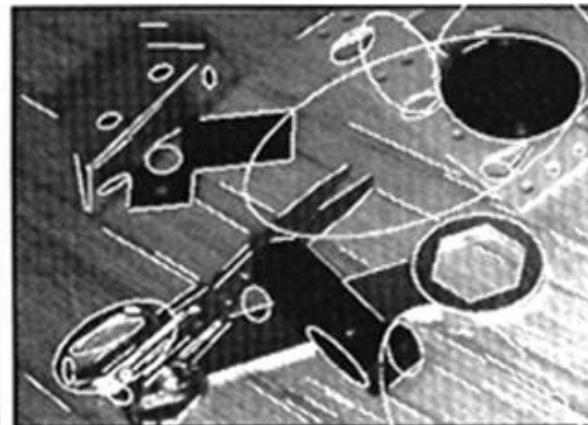
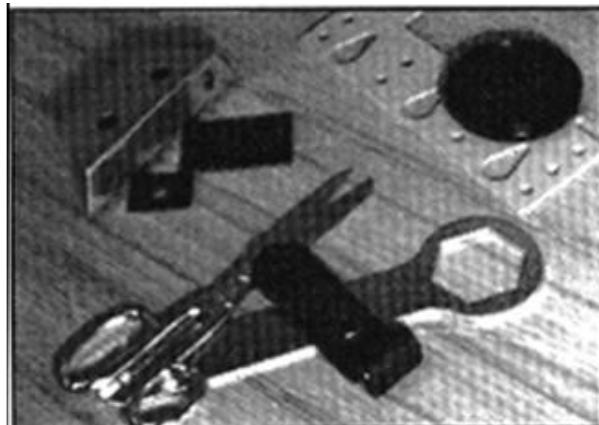
History of ideas in recognition

Example:
invariant to similarity transformations
computed from four points

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

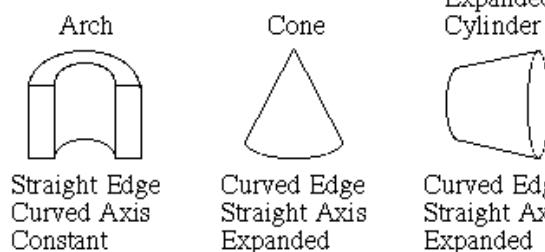
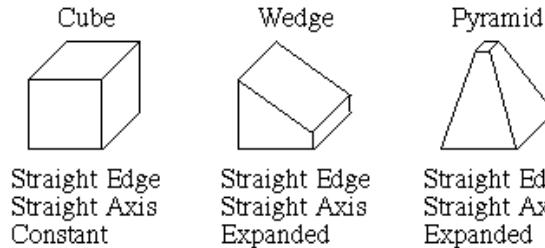


Projective invariants (Rothwell et al., 1992):

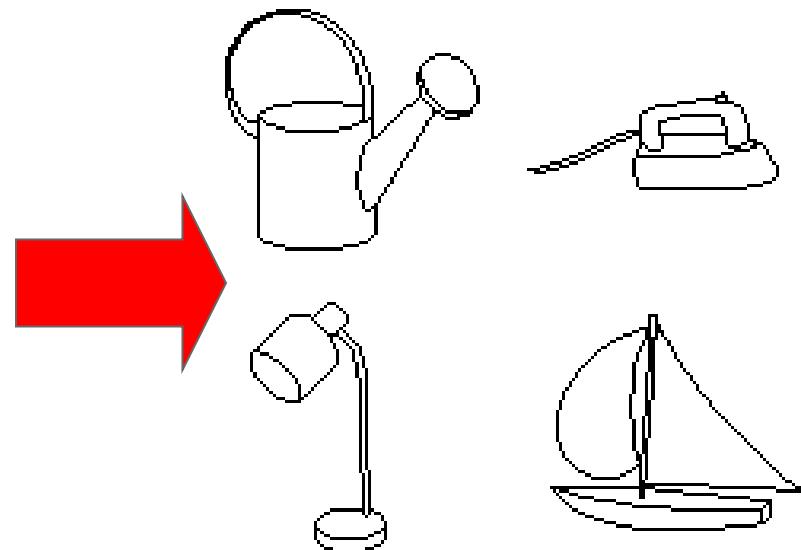


Recognition by components

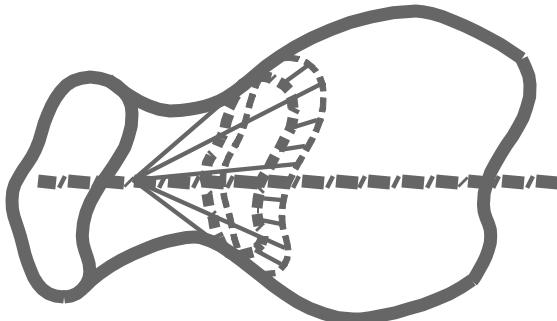
Primitives (geons) Biederman (1987)



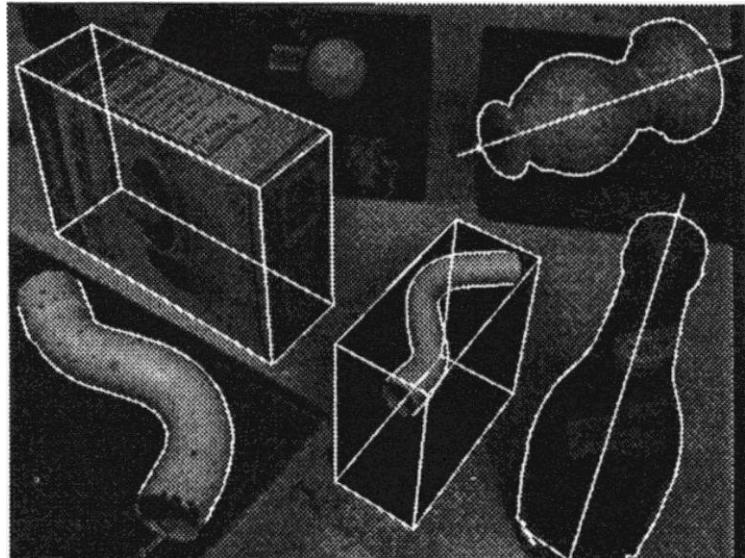
Objects



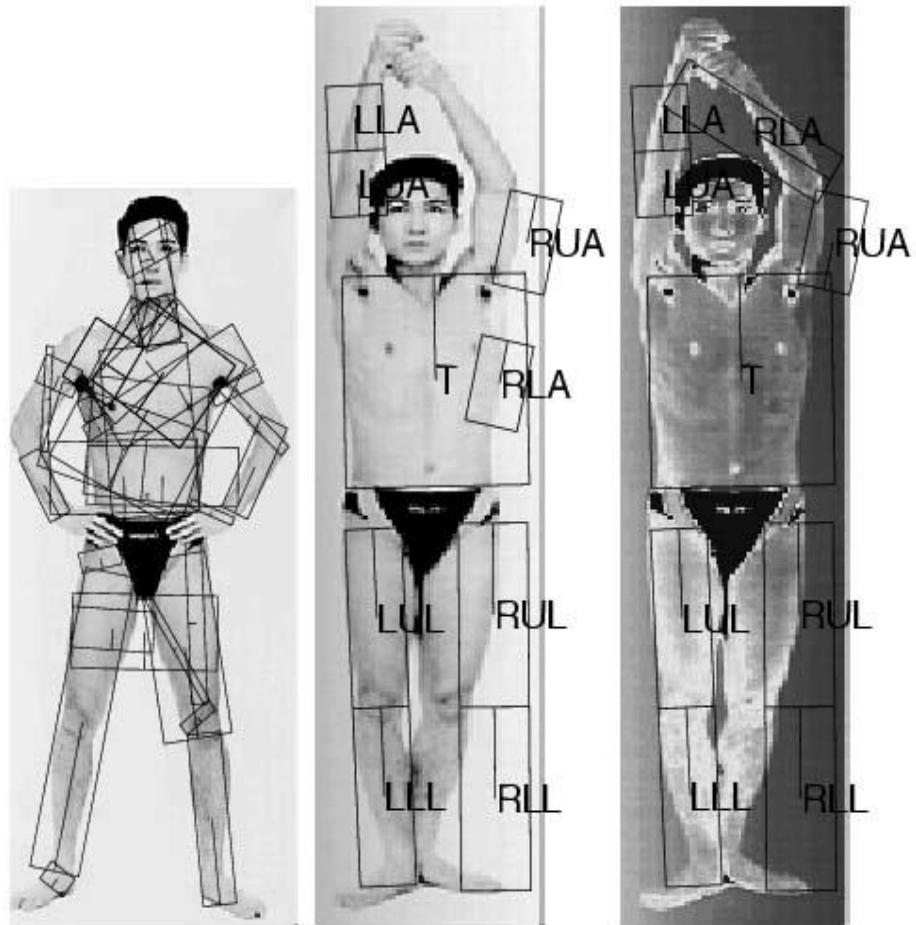
Recognition by components



Generalized cylinders
[Ponce et al. (1989)]



Zisserman et al. (1995)



Forsyth (2000)

History of ideas in recognition

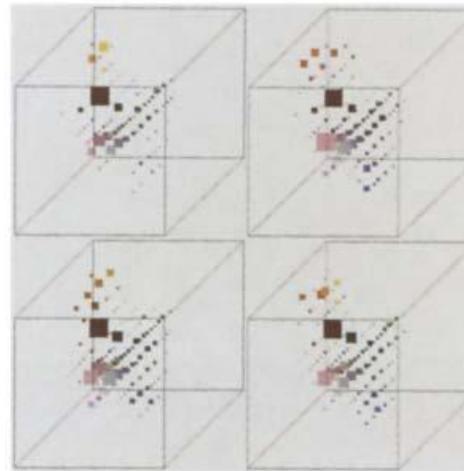
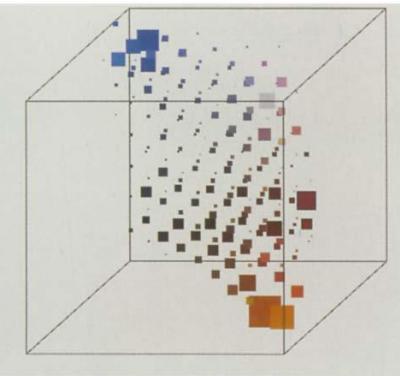
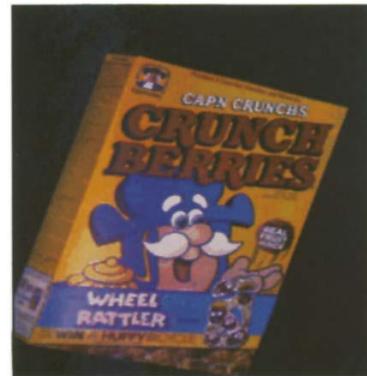
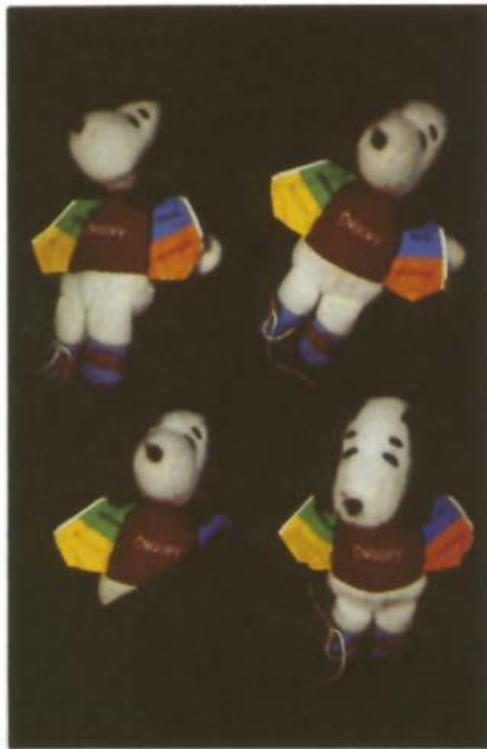
- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models

Empirical models of image variability

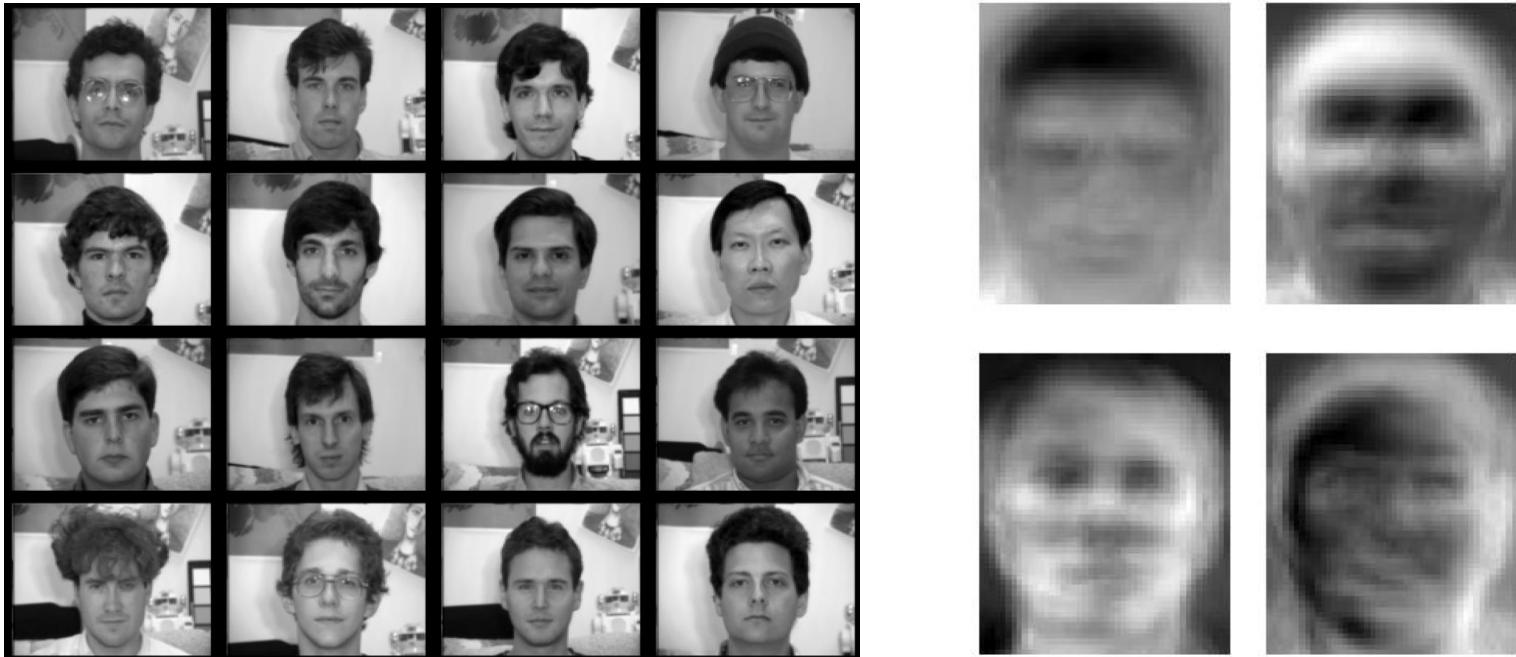
Appearance-based techniques

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

Color Histograms



Eigenfaces (Turk & Pentland, 1991)



| Experimental Condition | 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 |

History of ideas in recognition

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



Sliding window approaches



Turk and Pentland, 1991

Belhumeur, Hespanha, & Kriegman, 1997

Schneiderman & Kanade 2004

Viola and Jones, 2000



Schneiderman & Kanade, 2004

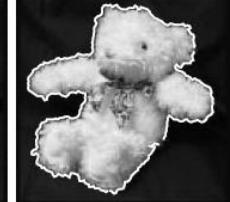
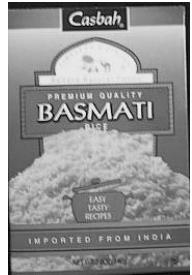
Argawal and Roth, 2002

Poggio et al. 1993

History of ideas in recognition

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

Local features for instance recognition



D. Lowe (1999, 2004)

Large-scale image search

Combining local features, indexing, and spatial constraints

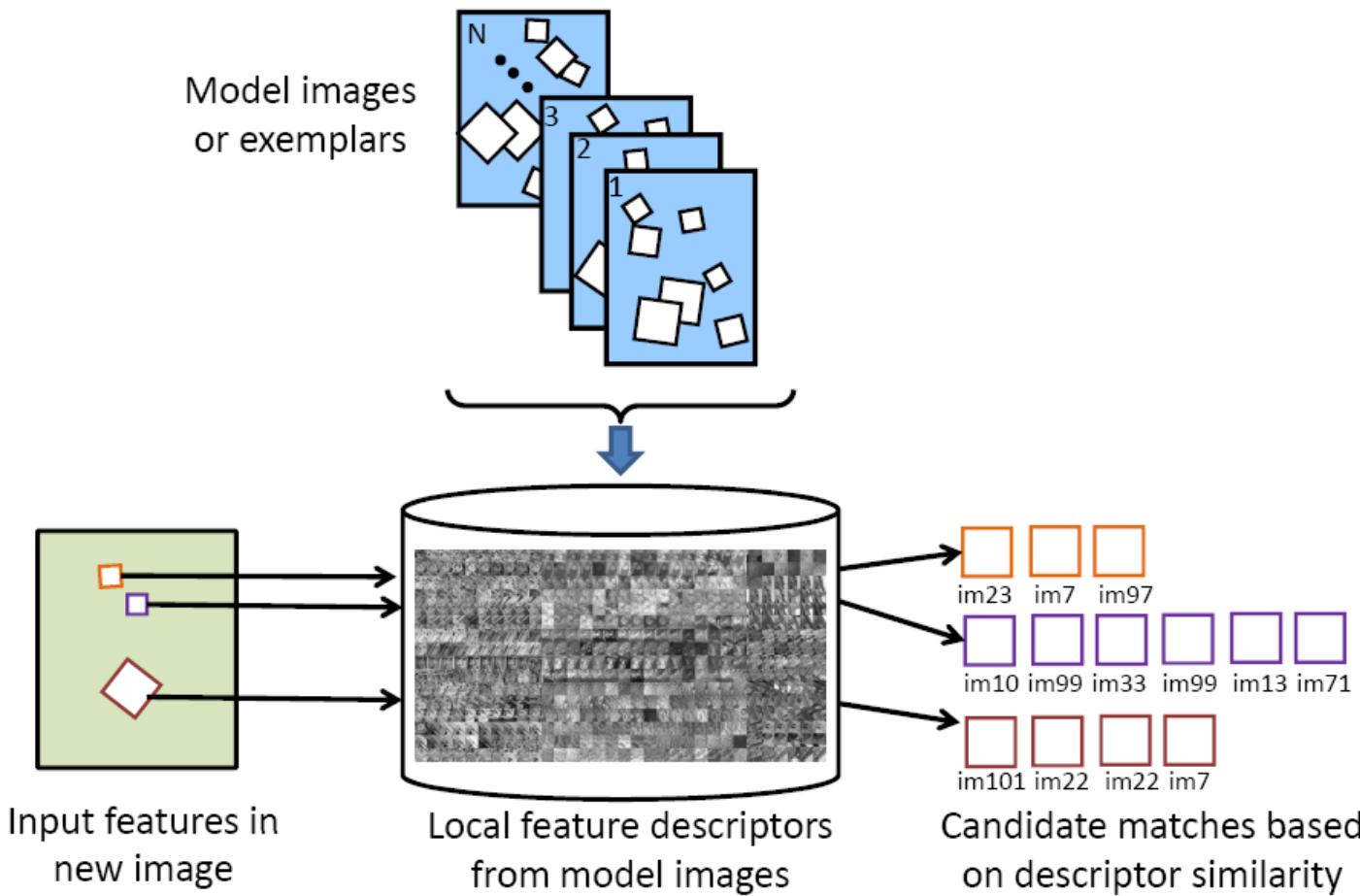
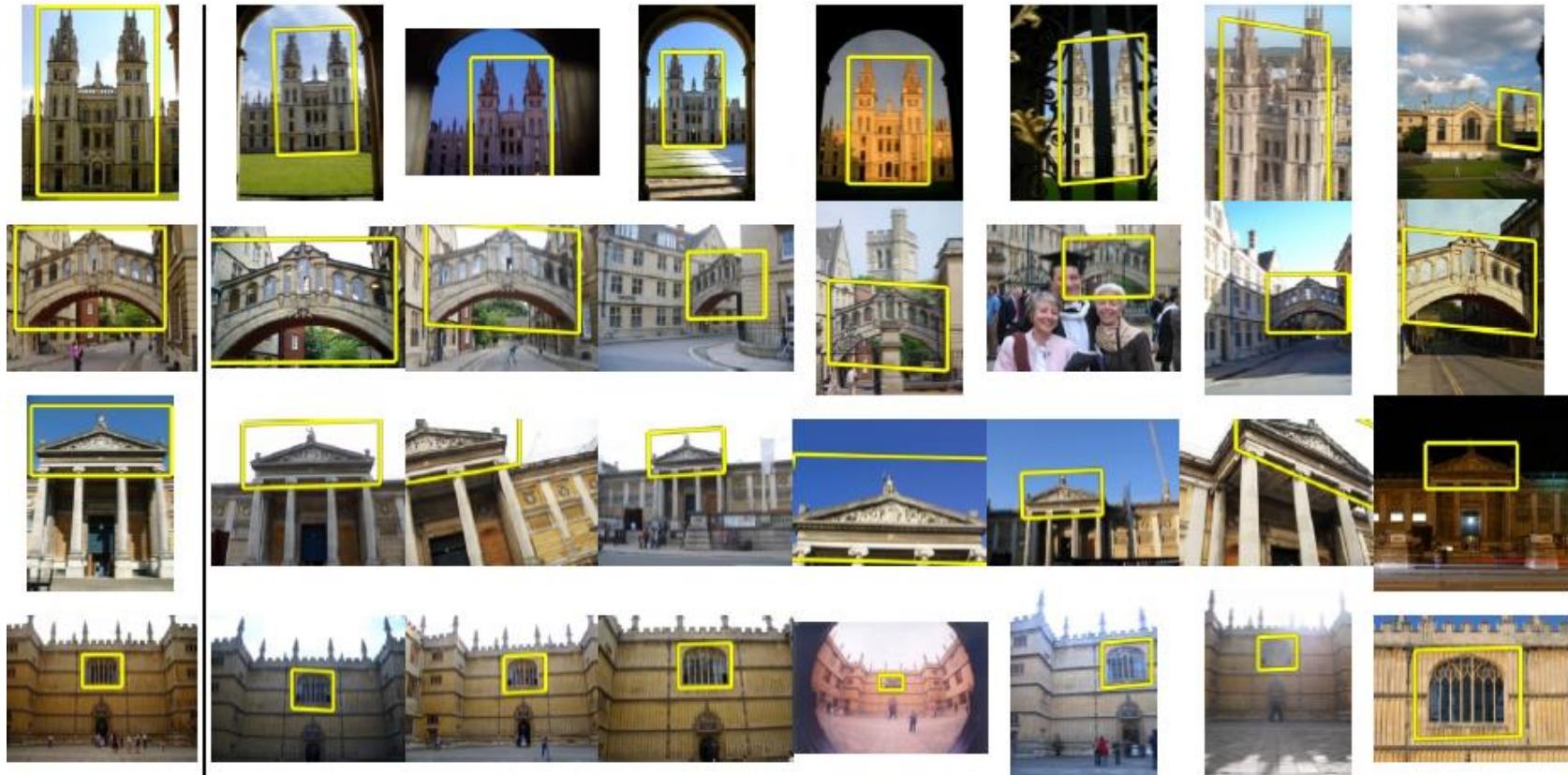


Image credit: K. Grauman and B. Leibe

Large-scale image search

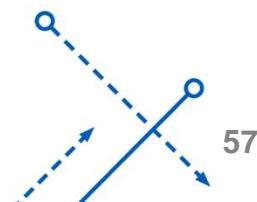
Combining local features, indexing, and spatial constraints



Philbin et al. '07

History of ideas in recognition

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

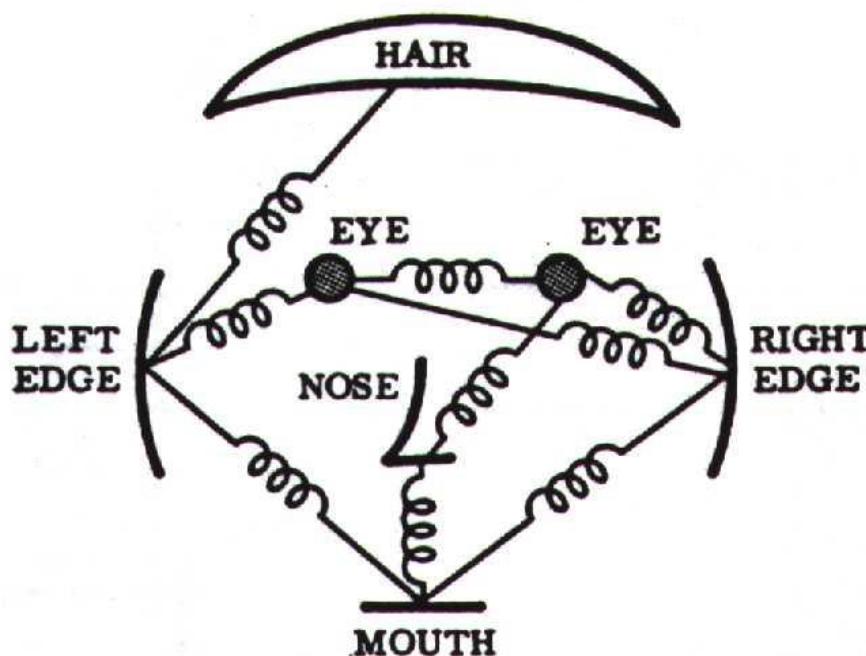


Figure from [Fischler & Elschlager 73]

Constellation models

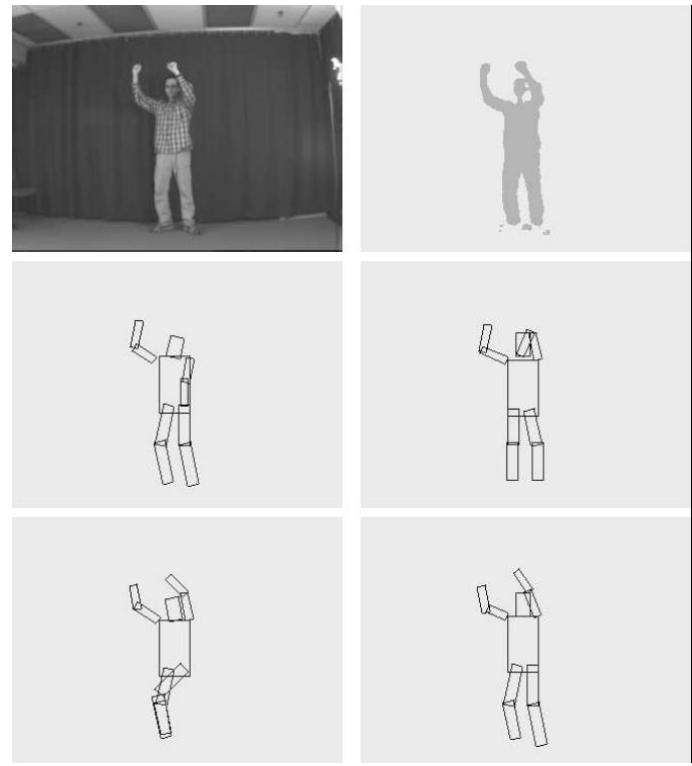
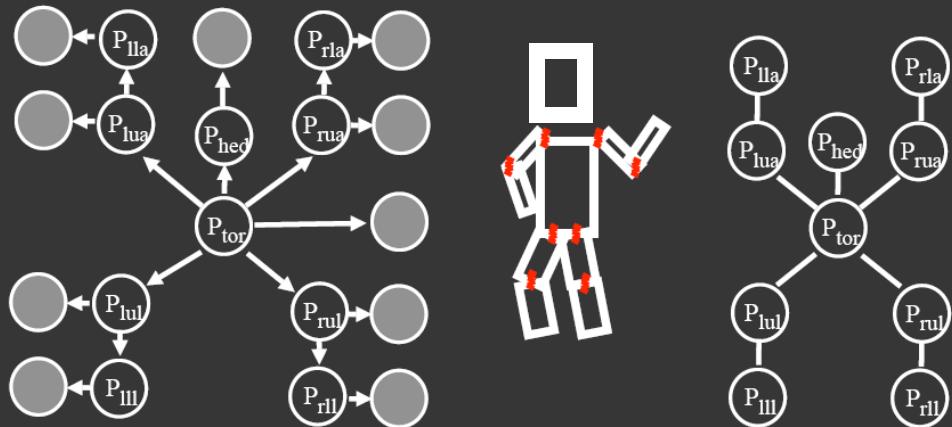


Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

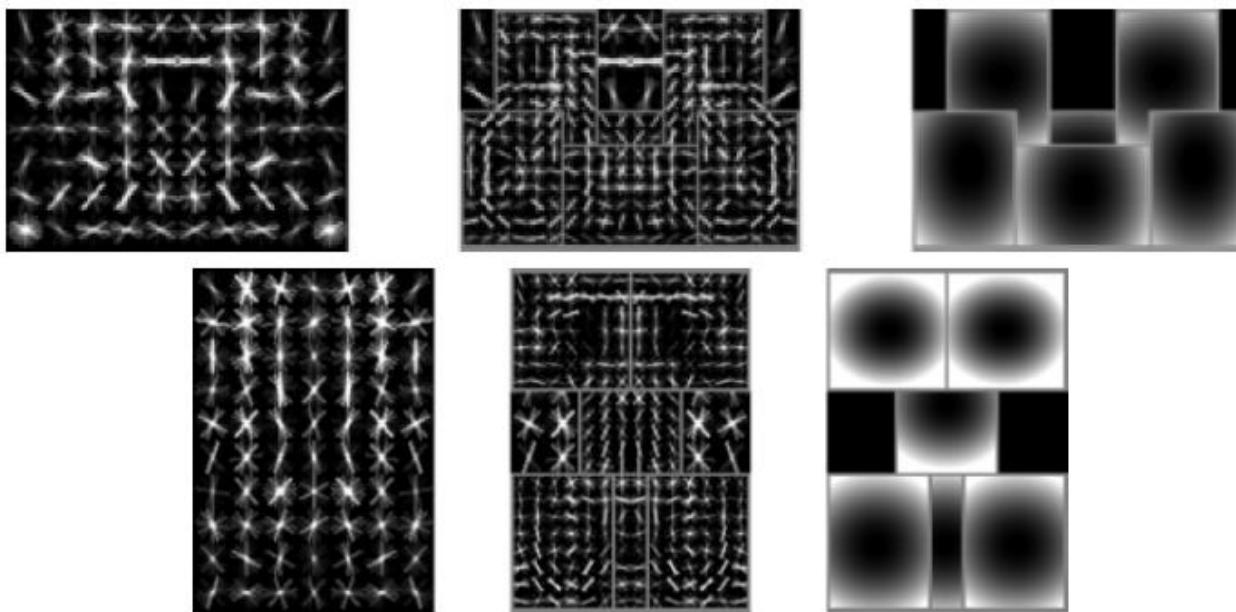
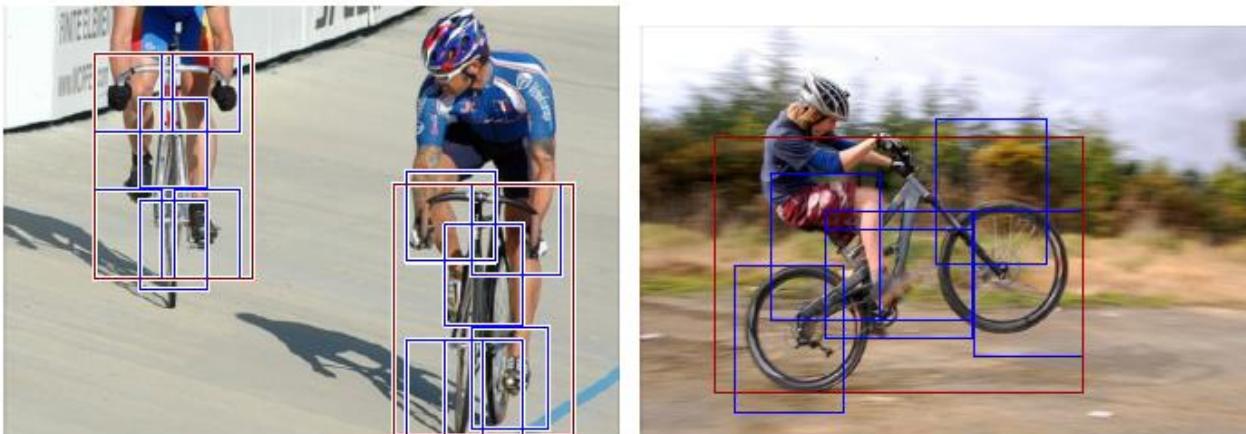
Representing people

Pictorial structure model

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



Discriminatively trained part-based models

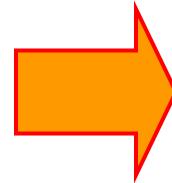


P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "[Object Detection with Discriminatively Trained Part-Based Models](#)," PAMI 2009

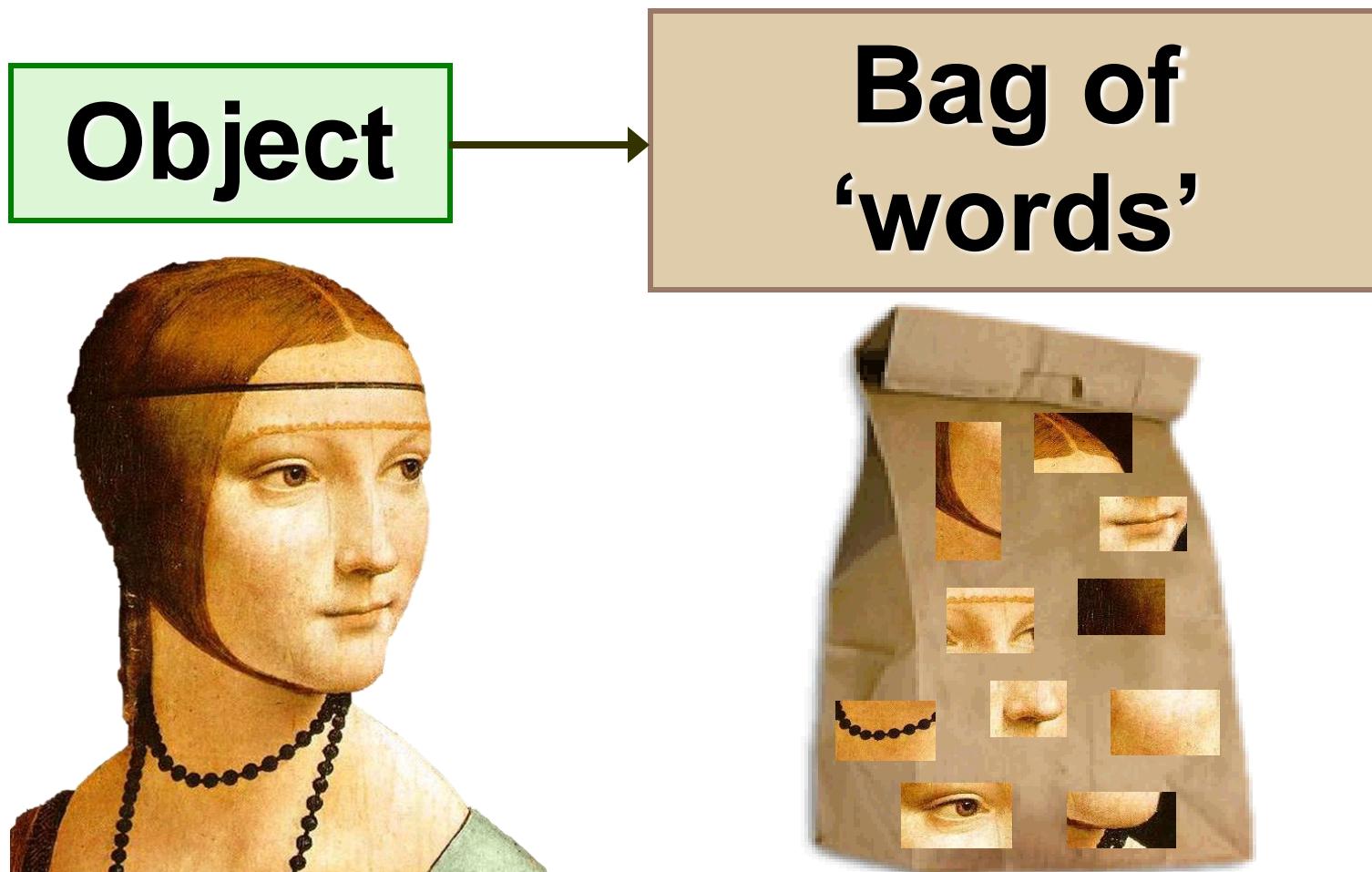
History of ideas in recognition

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

Bag-of-features models

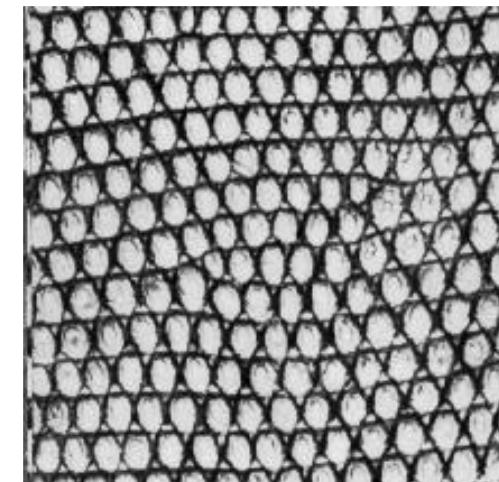
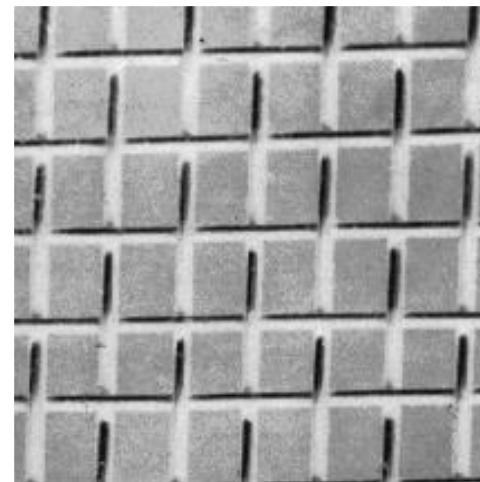
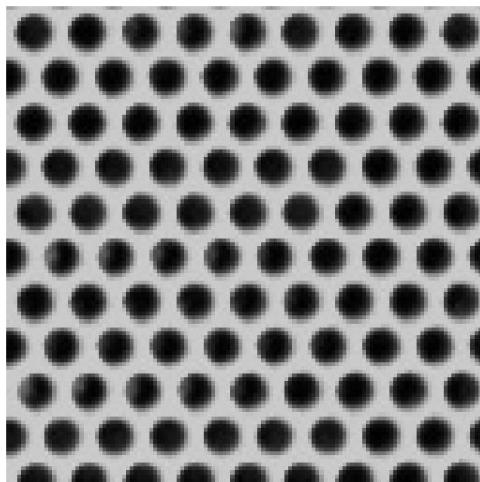


Bag-of-features models



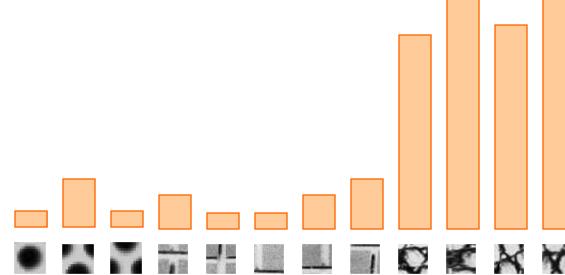
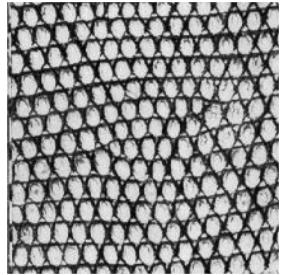
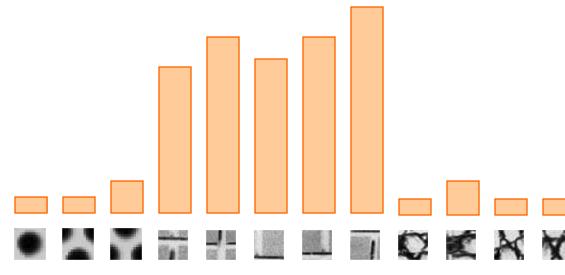
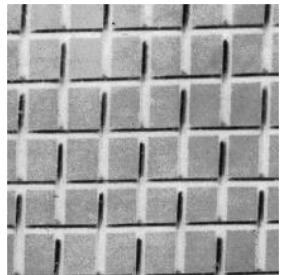
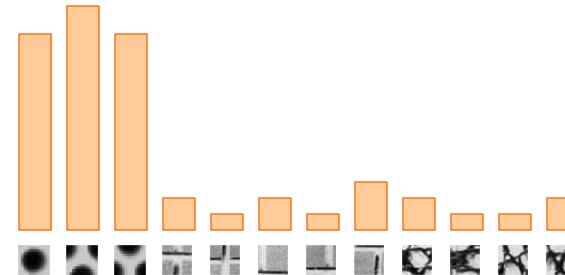
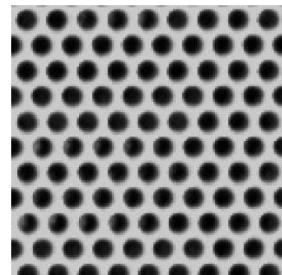
Origin 1: Texture recognition

- Texture is characterized by the **repetition of basic elements** or ***textons***
- For stochastic textures, the **identity** of the textons is more important than their **spatial arrangement**



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Origin 1: Texture recognition



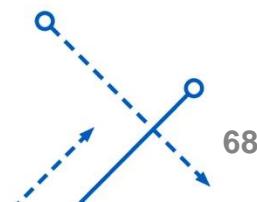
Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

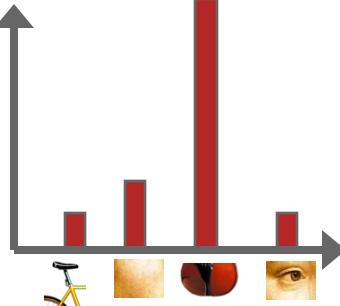
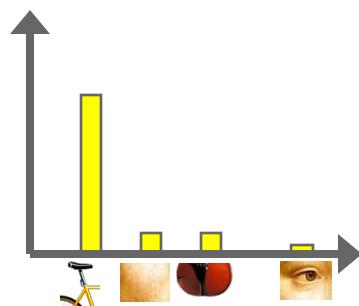
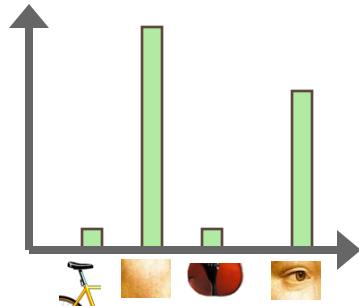
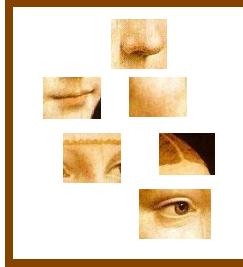


Bag-of-features steps

1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”

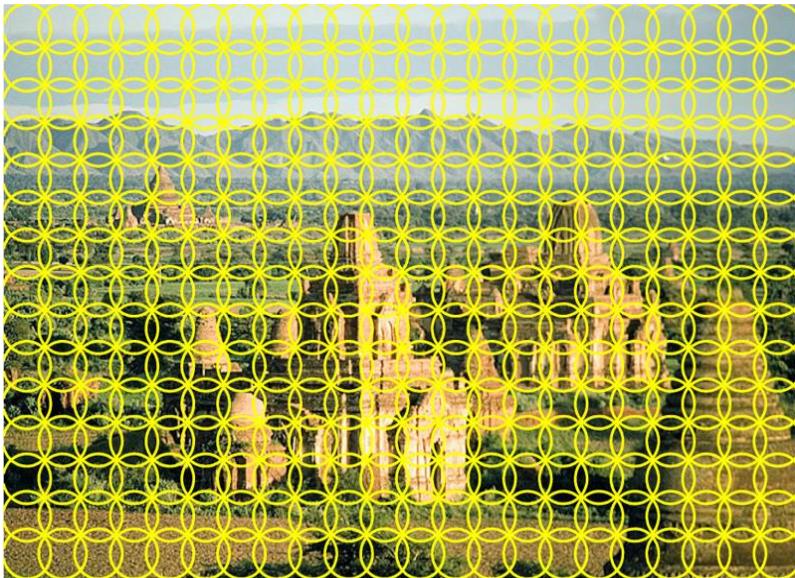


Bag-of-features steps



1. Feature extraction

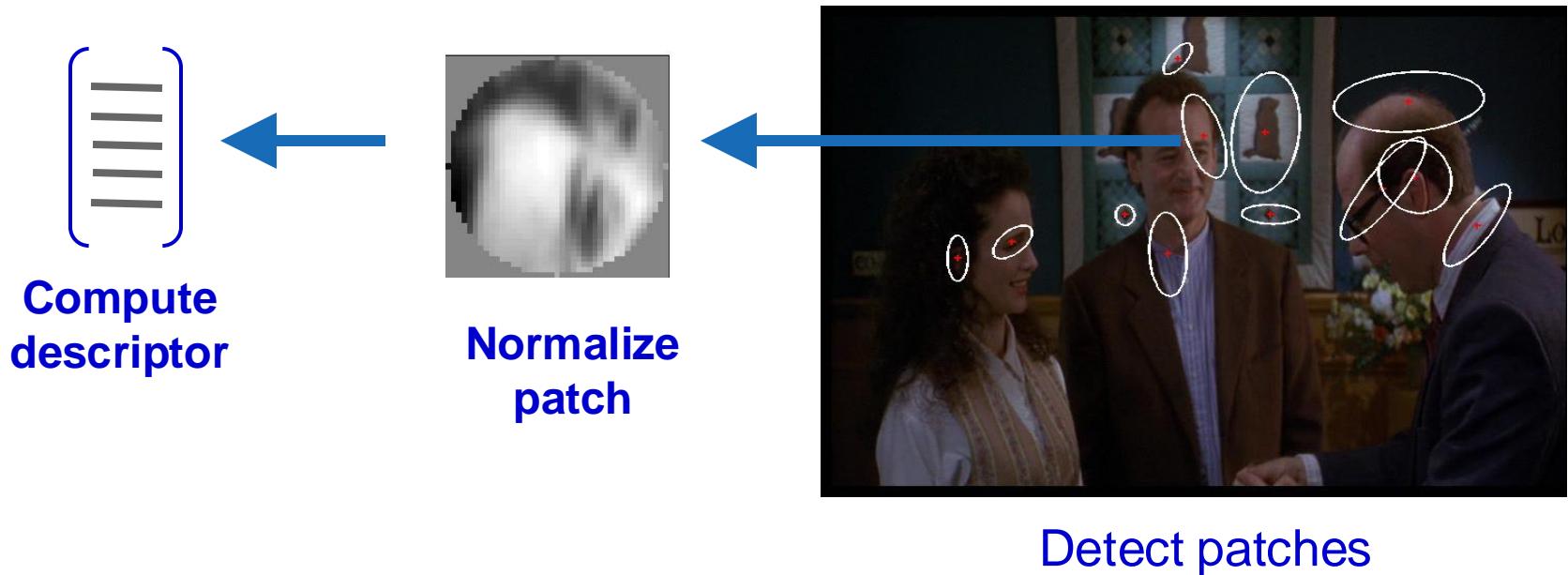
Regular grid



Interest regions

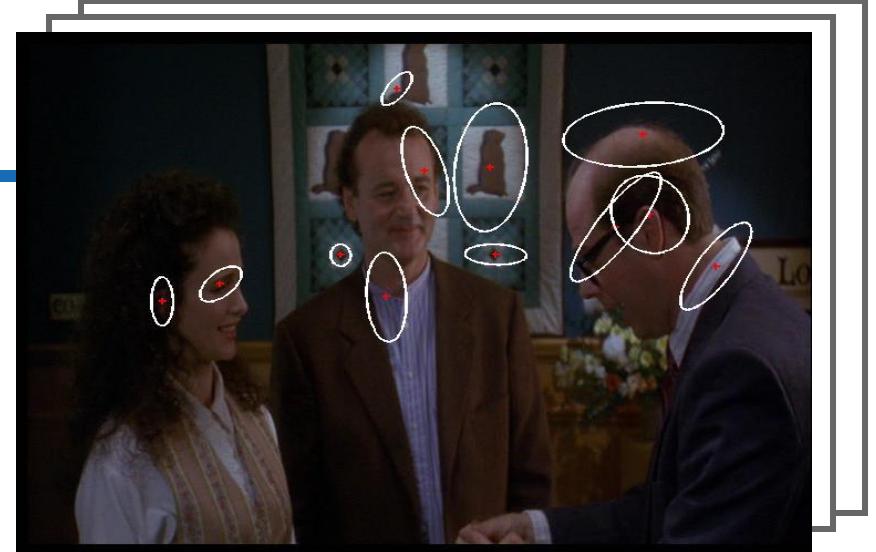
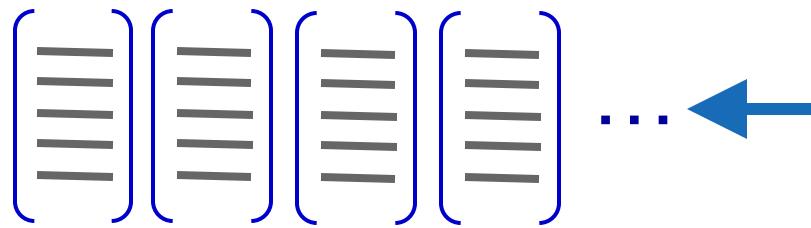


1. Feature extraction

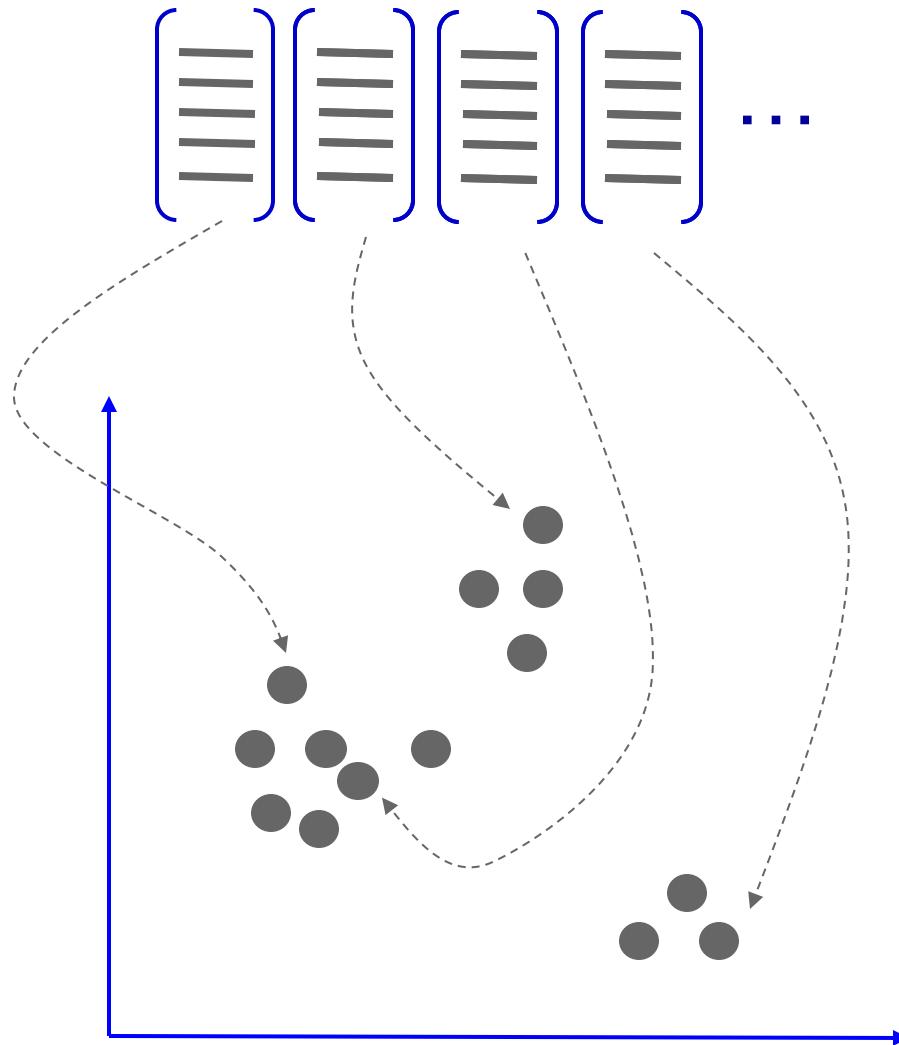


Slide credit: Josef Sivic 71

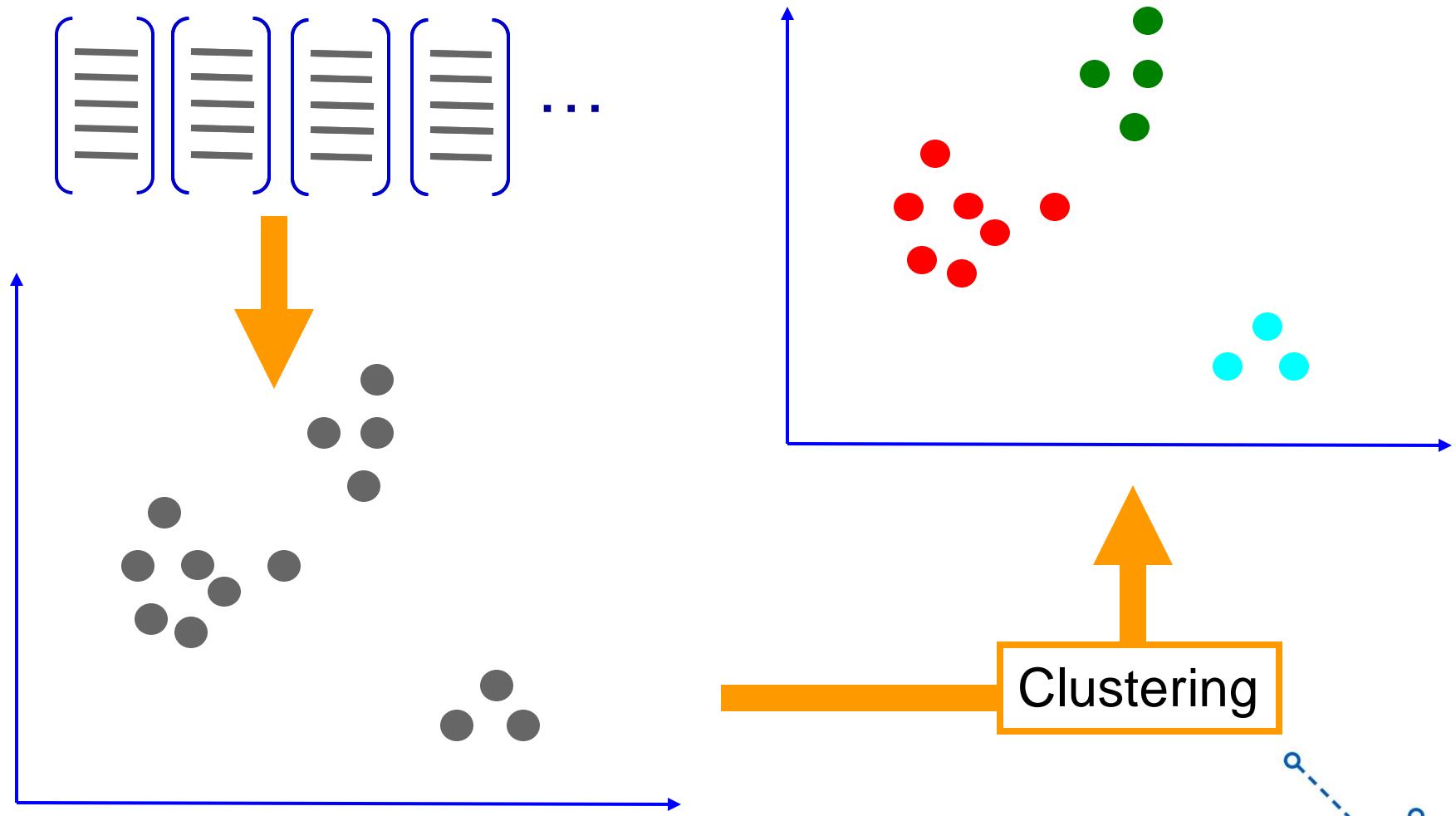
1. Feature extraction



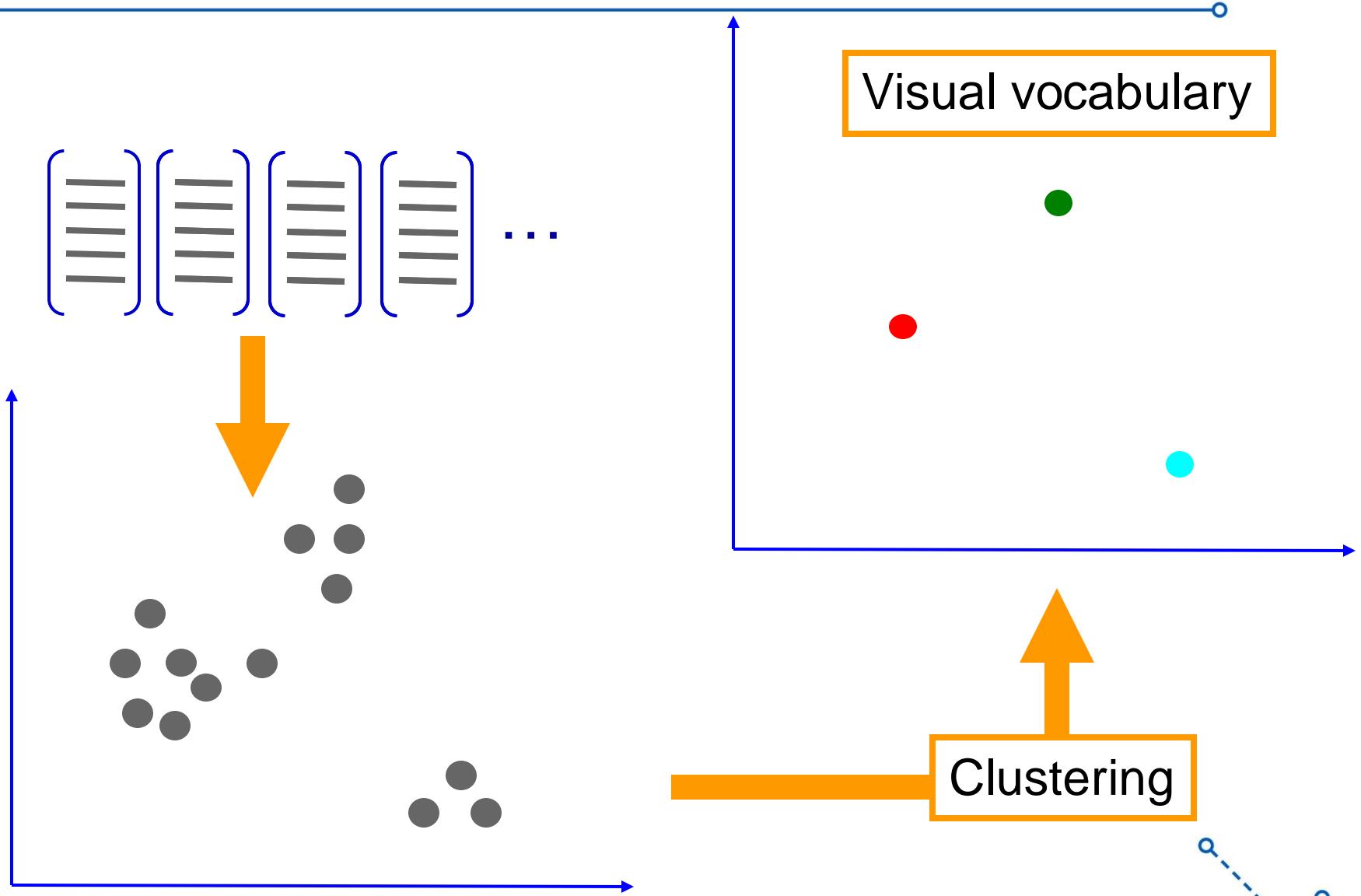
2. Learning the visual vocabulary



2. Learning the visual vocabulary

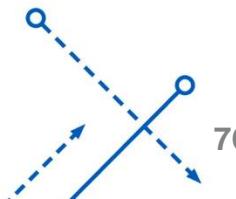


2. Learning the visual vocabulary

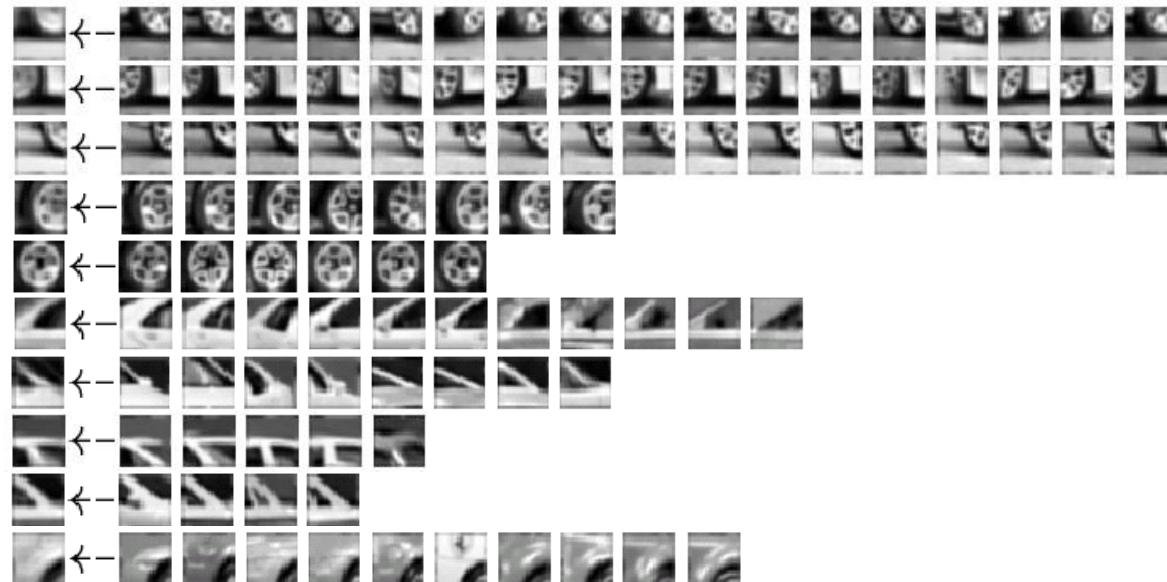
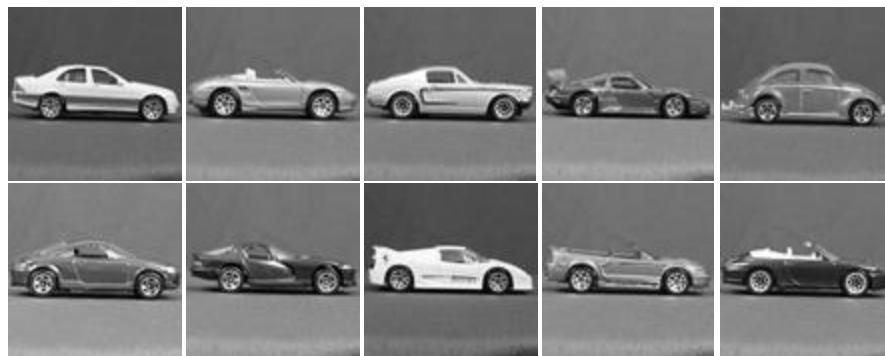


Clustering and vector quantization

- Clustering: learning a visual vocabulary or codebook
 - Unsupervised learning process
 - Each cluster center (e.g., k-means) becomes a codevector
 - Codebook can be learned on a separate training set
 - If the training set is sufficiently representative, the codebook will be “universal”
- The codebook is used for quantizing features
 - A *vector quantizer* takes a feature vector and maps it to the index of the nearest codevector in a codebook
 - Codebook = visual vocabulary
 - Codevector = visual word

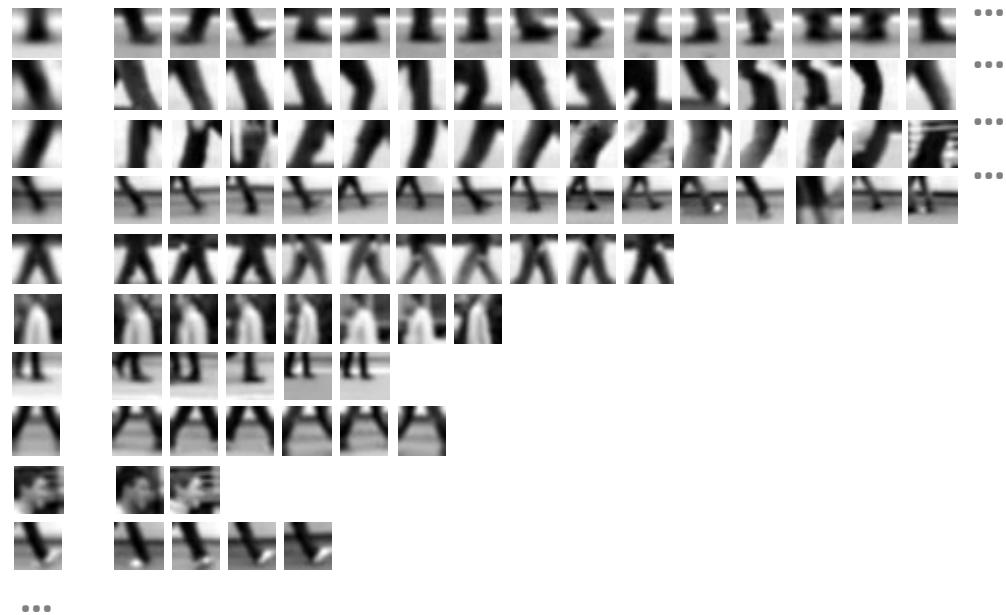
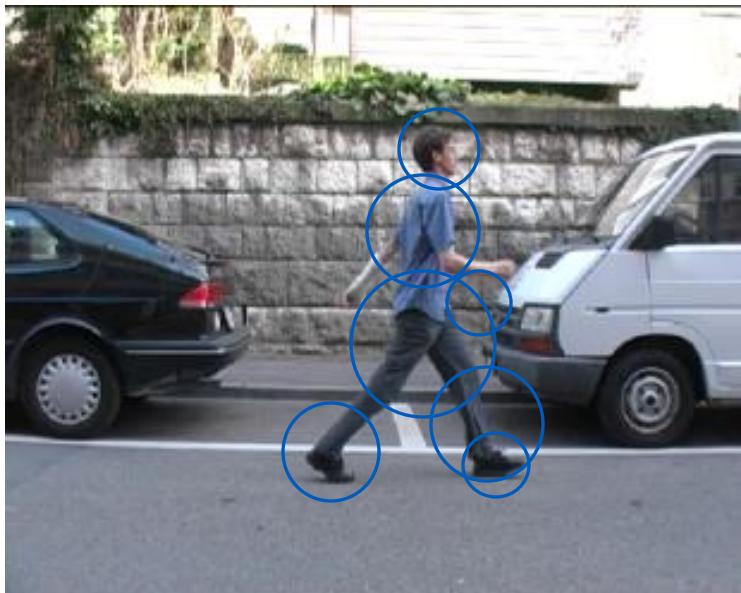


Example Codebook



Appearance codebook

Another codebook



Appearance codebook