Computer Vision Course
Lecture 09

Recognition 02

Ceyhun Burak Akgül, PhD
cba-research.com

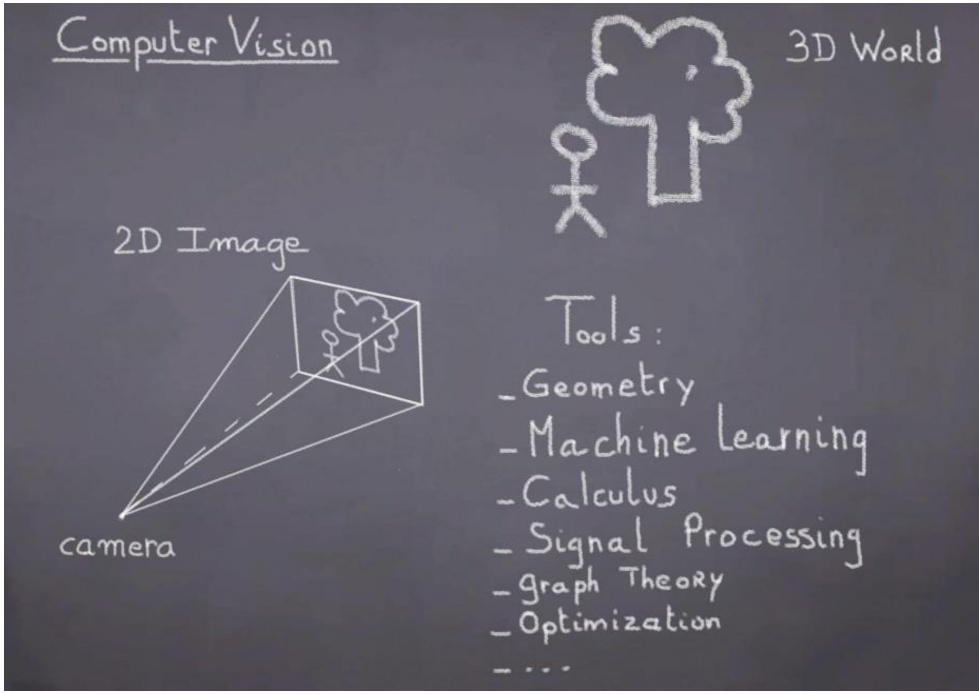
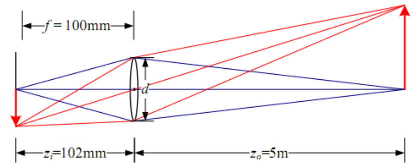


Photo credit: Olivier Teboul
vision.mas.ecp.fr/Personnel/teboul

Spring 2015
Last updated 06/05/2015

These slides have been adapted from James Hays's 2014 Computer Vision course slides at Brown University.

Course Outline



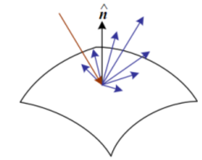
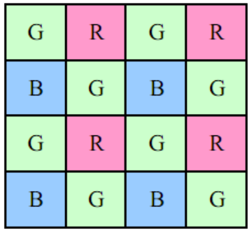


Image Formation and Processing

Light, Shape and Color

The Pin-hole Camera Model, The Digital Camera

Linear filtering, Template Matching, Image Pyramids

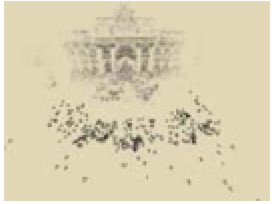


Feature Detection and Matching

Edge Detection, Interest Points: Corners and Blobs

Local Image Descriptors

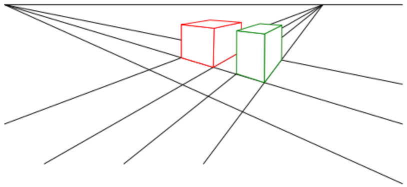
Feature Matching and Hough Transform



Multiple Views and Motion

Geometric Transformations, Camera Calibration

Feature Tracking , Stereo Vision



Segmentation and Grouping

Segmentation by Clustering, Region Merging and Growing

Advanced Methods Overview: Active Contours, Level-Sets, Graph-Theoretic Methods

Detection and Recognition

Problems and Architectures Overview

Statistical Classifiers, Bag-of-Words Model, **Detection by Sliding Windows**

History of Ideas in Recognition

Visual Recognition Problems – *recap*

Object Instance Recognition

Recognize different instances of the same object (e.g., a product package, a face, a specific mug) given an image that tightly contains a single object

Object Category Recognition

Recognize different examples of the same object category (e.g., car, airplane, flower) given an image that tightly contains a single object

Object Detection and Localization

Do the above (instance or category) on an image containing the object at arbitrary position and scale

Image Classification

Classify an image based on its content (indoor/outdoor, nature/urban, sunny/cloudy/rainy, Paris/Istanbul/…, etc.)

Scene Understanding

Tell what is going in the image, e.g., "a car running on the high way at sunset, it's summer time, …"

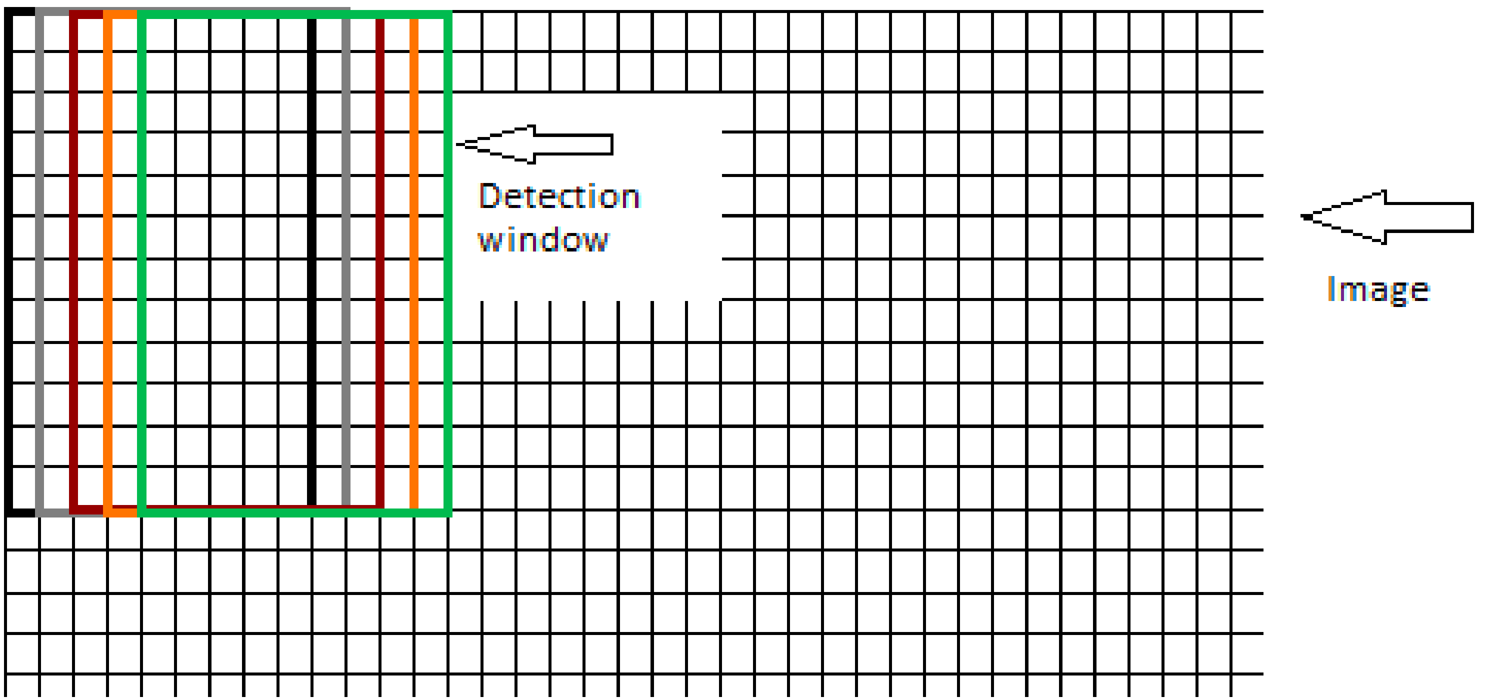
Architectures

- Aligned Representations
- Voting Schemes: Generalized Hough Transform
- Bag-of-Words Model
- Detection by Sliding Windows
- Parts-based Models

Architectures – *recap*

- Aligned Representations – *last week*
- Voting Schemes: Generalized Hough Transform – *seen*
- Bag-of-Words Model – *last week*
- **Detection by Sliding Windows**
- Parts-based Models – *not in this class*

Detection by Sliding Windows – 1/4



Slide windows over the image with a stride parameter *s* (here s = 1 pixel)

Detection by Sliding Windows – 2/4

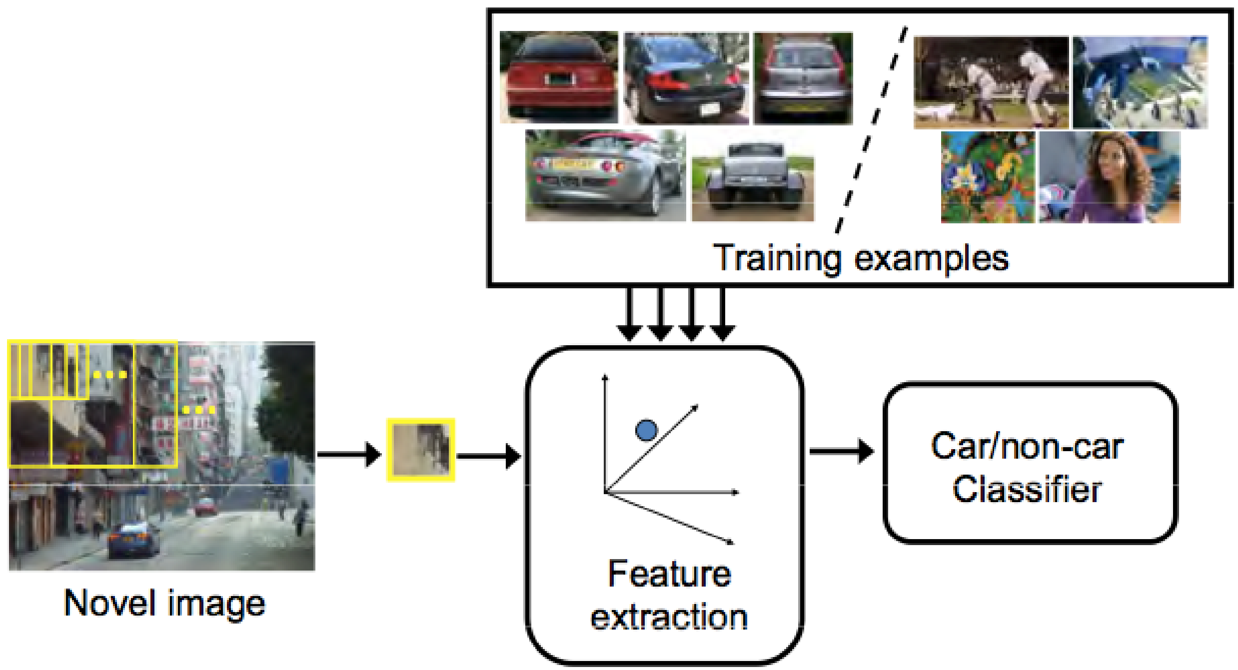
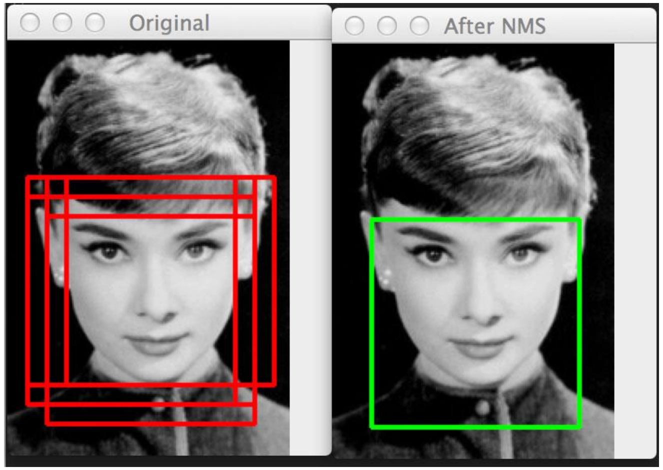


Figure 9.1: Main components of a sliding window detector. To learn from the images, some feature representation must be selected. Labeled examples (positive exemplars, or both negative and positive exemplars) are used to train a classifier that computes how likely it is that a given window contains the object category of interest. Given a novel image, the features from each of its sub-windows at multiple scales are extracted, and then tested by the classifier.

Detection by Sliding Windows – 3/4

Figure 9.2: Non-maximum suppression is a useful post-processing step to prune out nearby detections.



Detection by Sliding Windows – 4/4

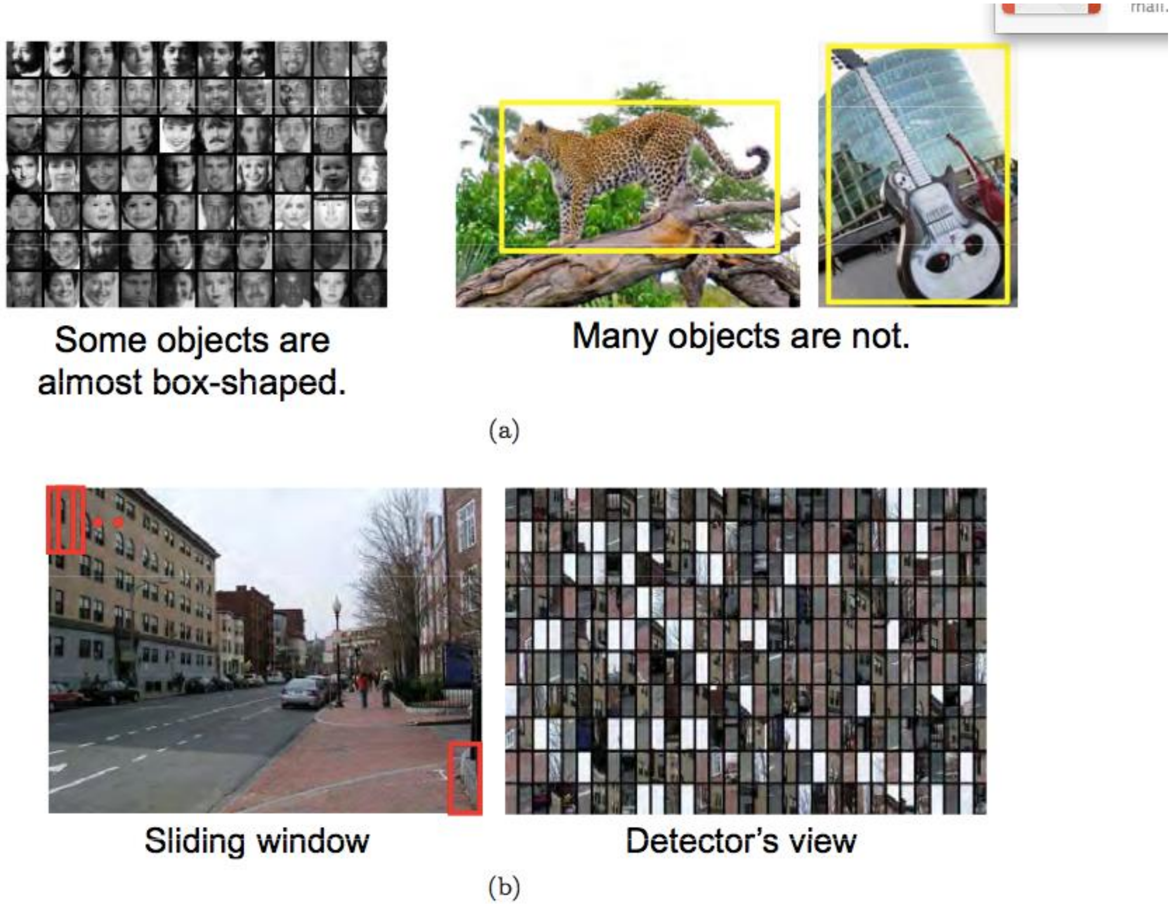
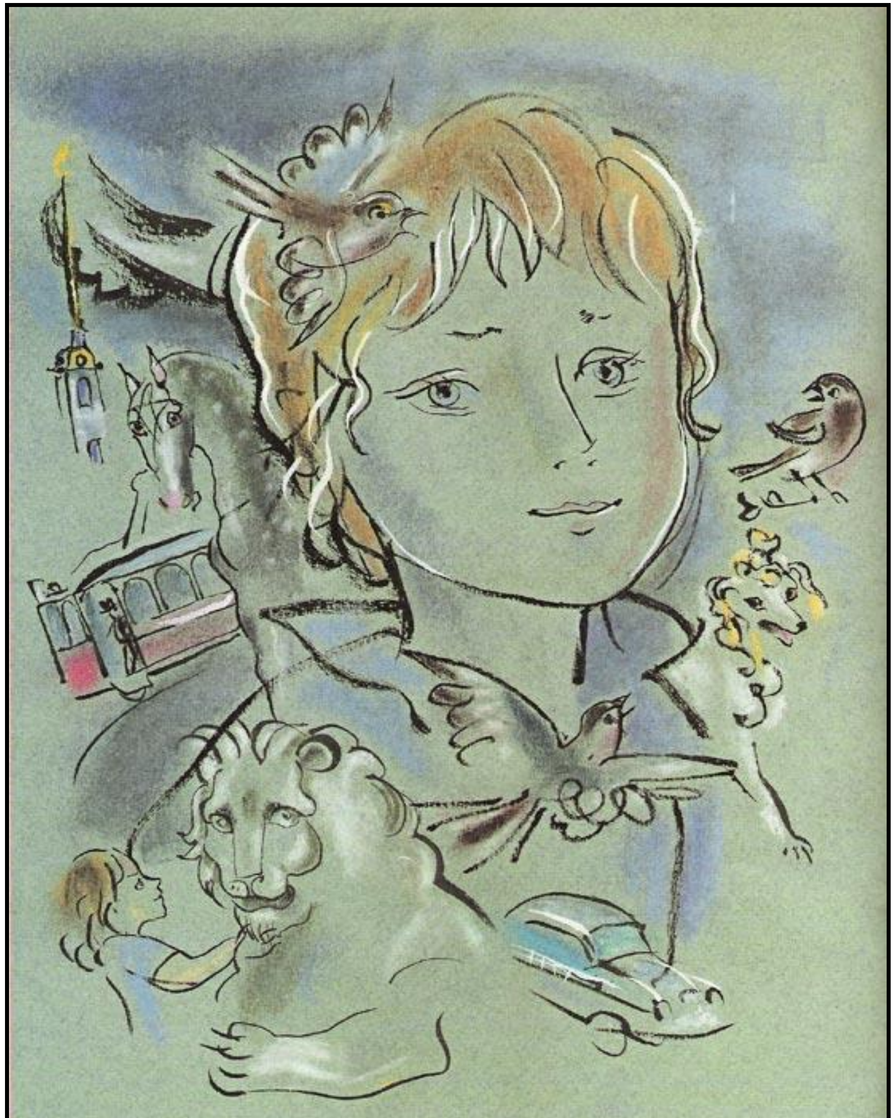


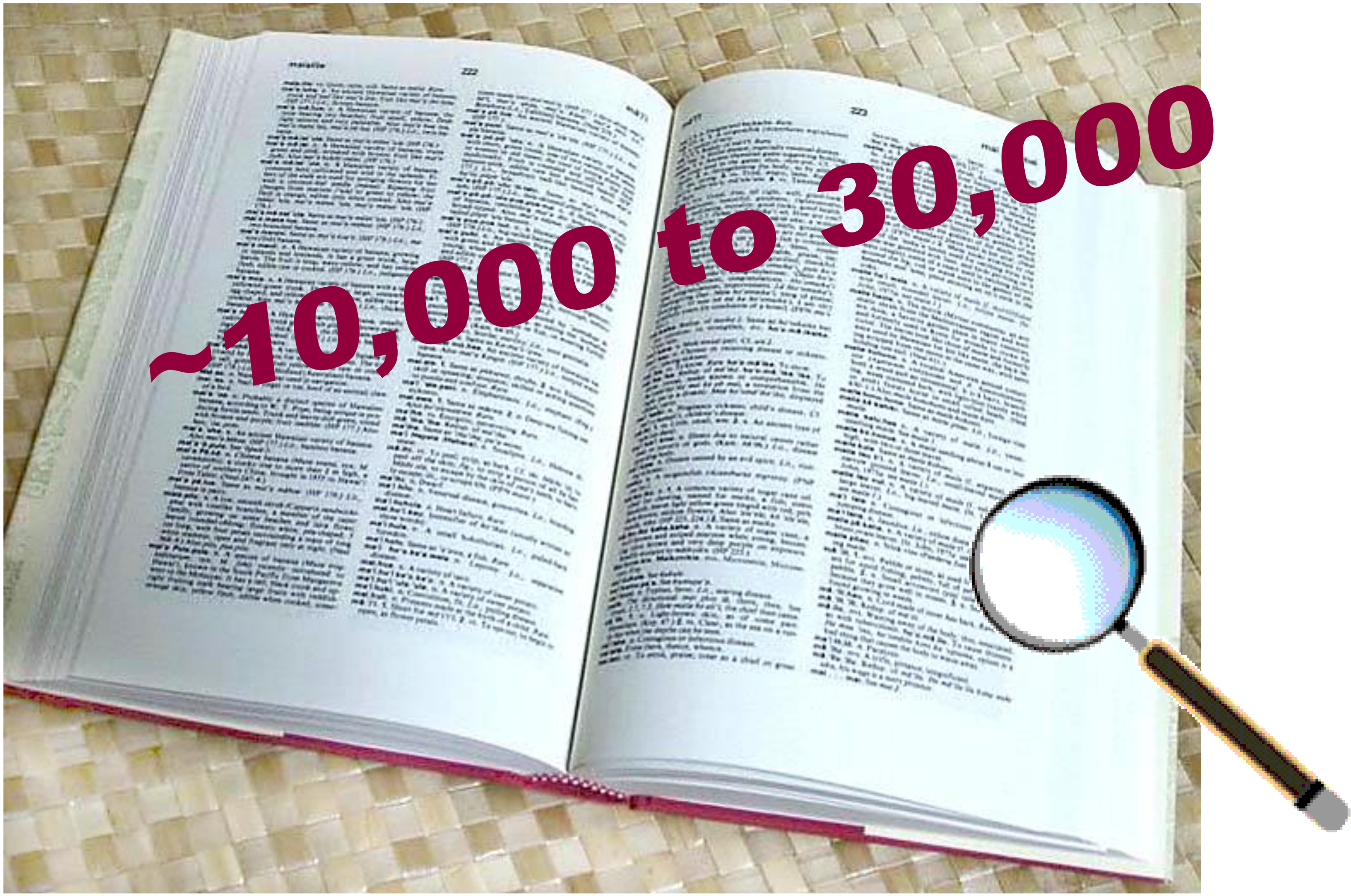
Figure 9.3: Sliding window detectors have noted limitations. Not all object categories are captured well by a consistent and box-shaped appearance pattern (a), and considering windows in isolation misses out on a great deal of information provided by the scene (b). FACE IMAGES FROM Viola 2001, (b) IS FROM DEREK HOIEM'S SLIDES.

Recognition: Overview and History



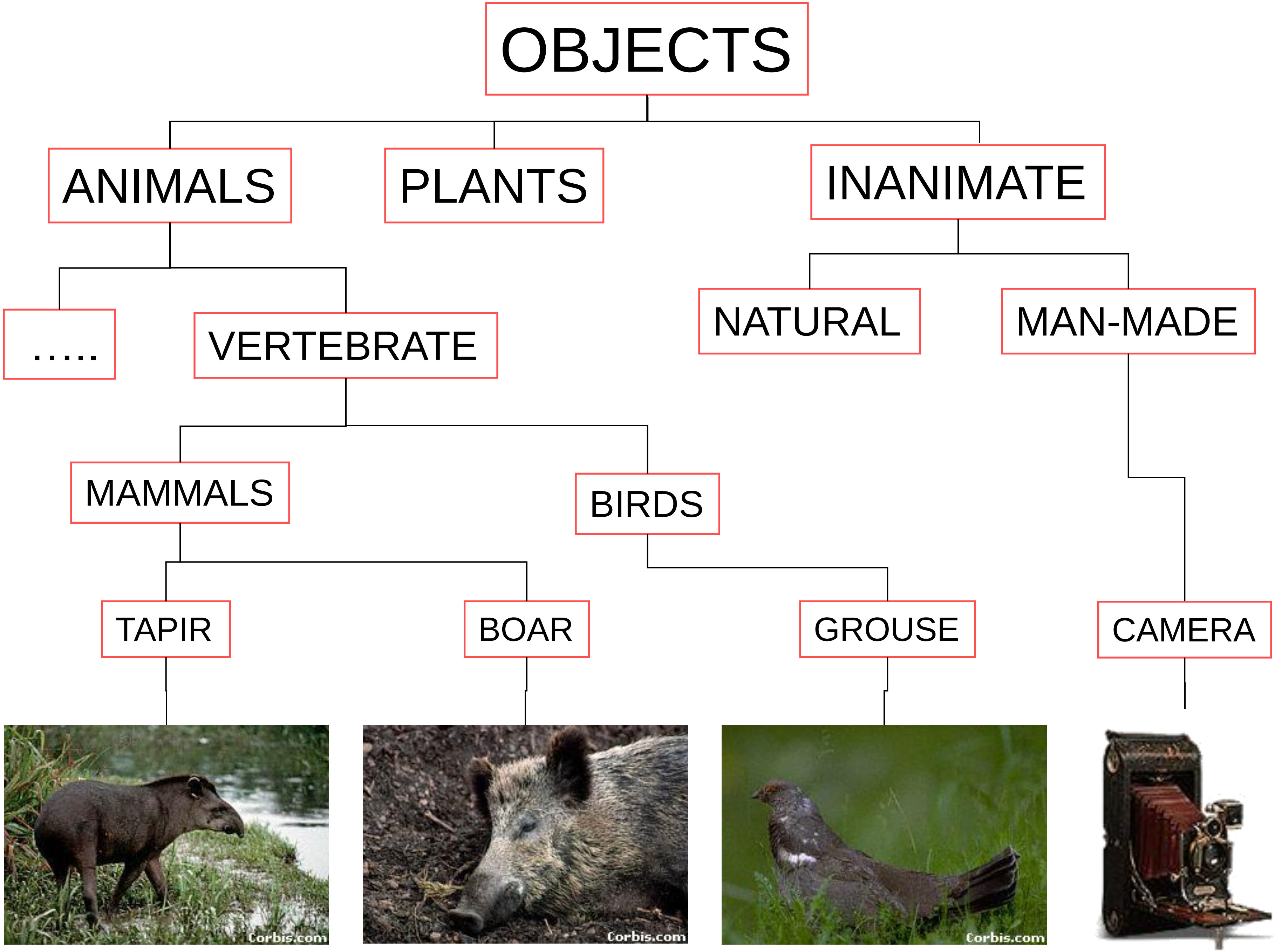
Slides from Lana Lazebnik, Fei-Fei Li, Rob Fergus, Antonio Torralba, and Jean Ponce

How many visual object categories are there?



Biederman 1987


~10,000 to 30,000


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

Specific recognition tasks

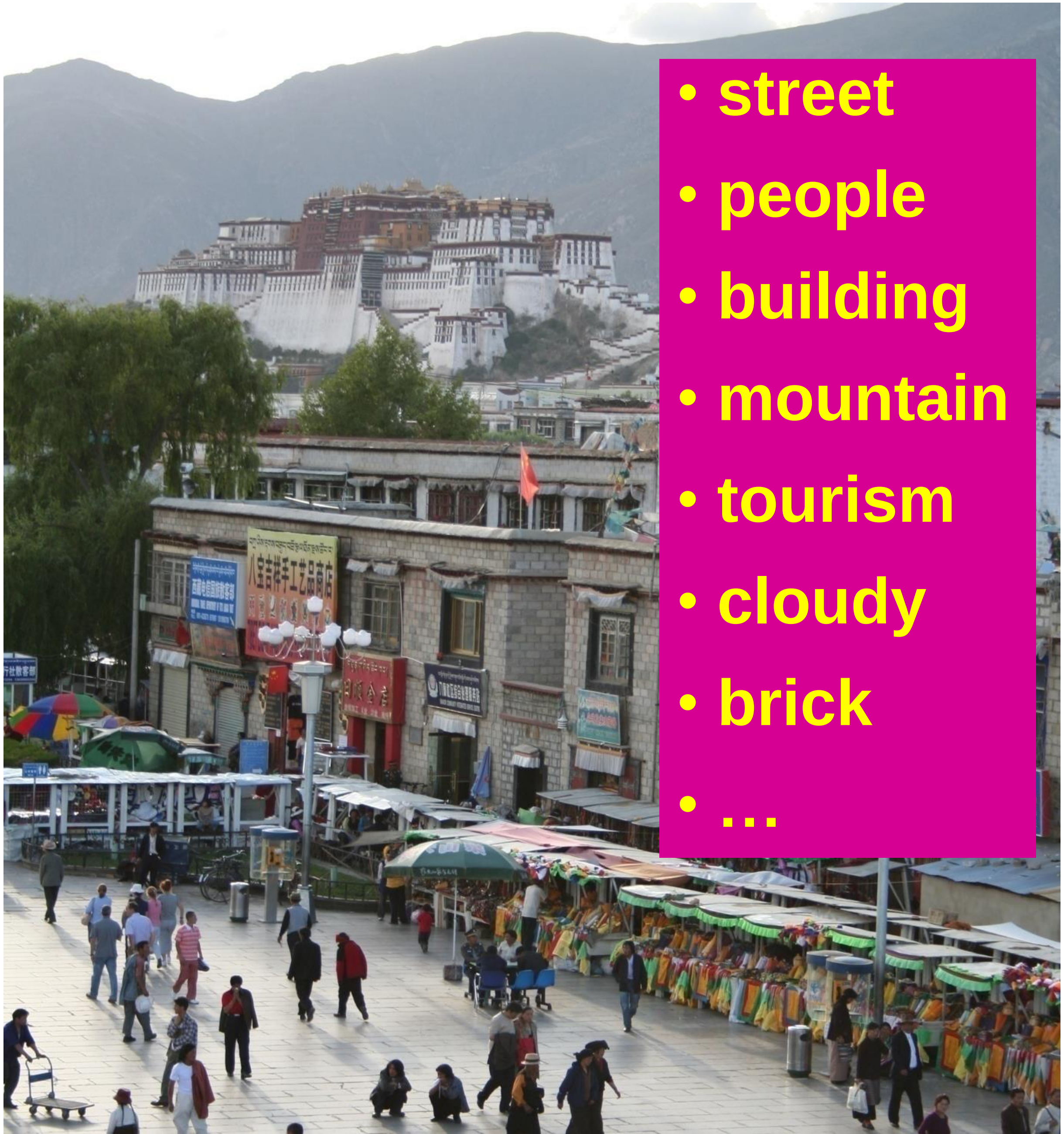
Svetlana Lazebnik

Scene categorization or classification

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

Svetlana Lazebnik

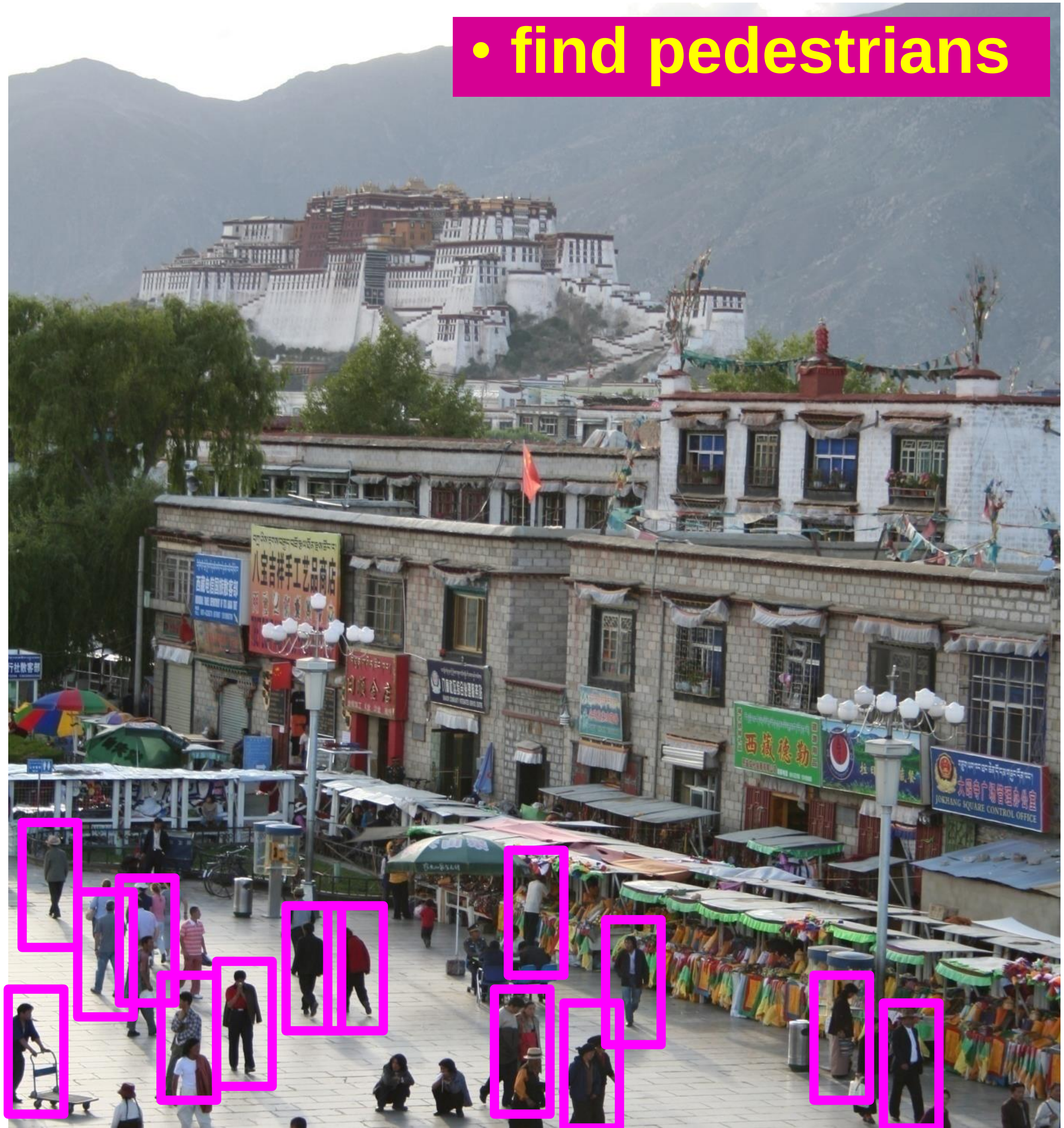
Image annotation / tagging / attributes



- street
- people
- building
- mountain
- tourism
- cloudy
- brick
- ...

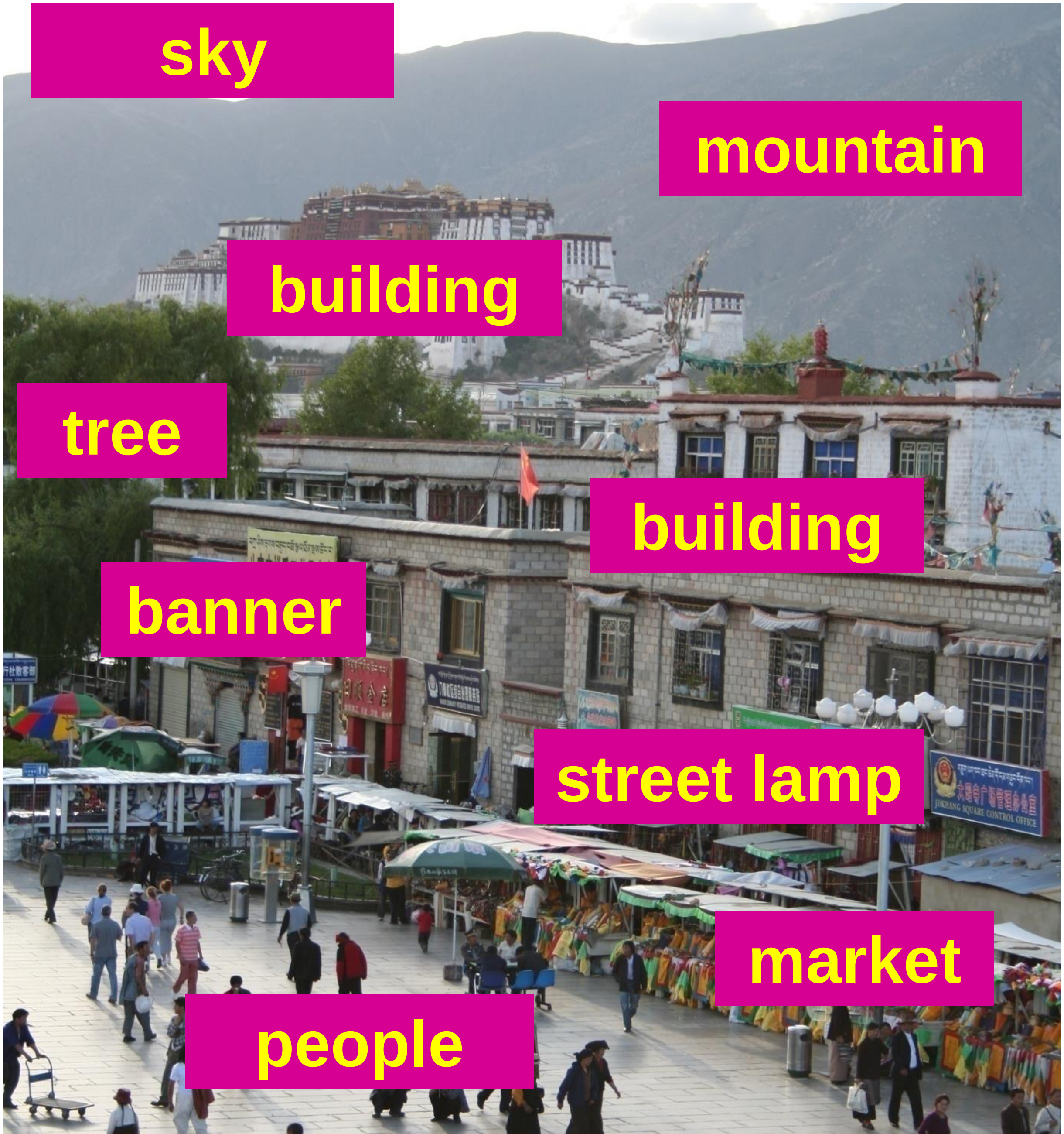
Svetlana Lazebnik

Object detection



Svetlana Lazebnik

Image parsing / semantic segmentation

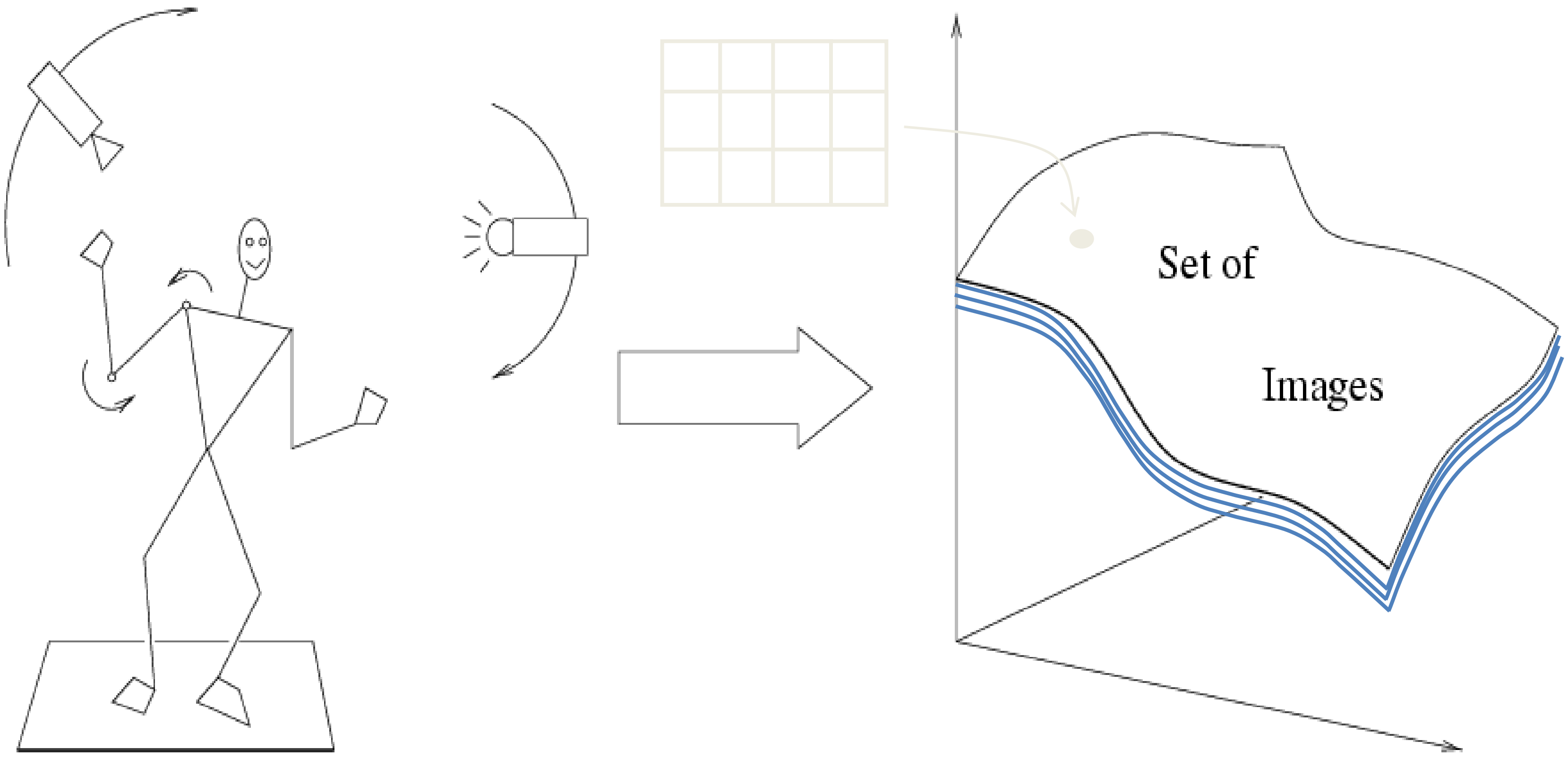


Svetlana Lazebnik

Scene understanding?

Svetlana Lazebnik

Recognition is all about modeling variability

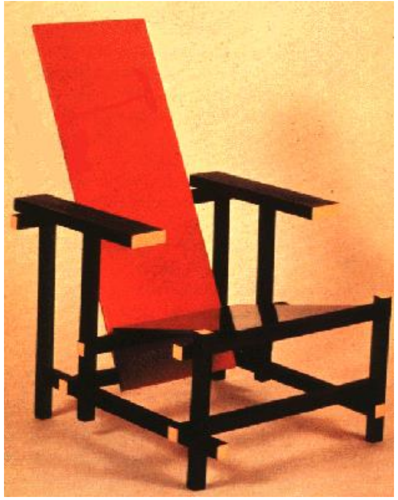


Variability: Camera position
Illumination
Shape parameters

Within-class variations?

Svetlana Lazebnik

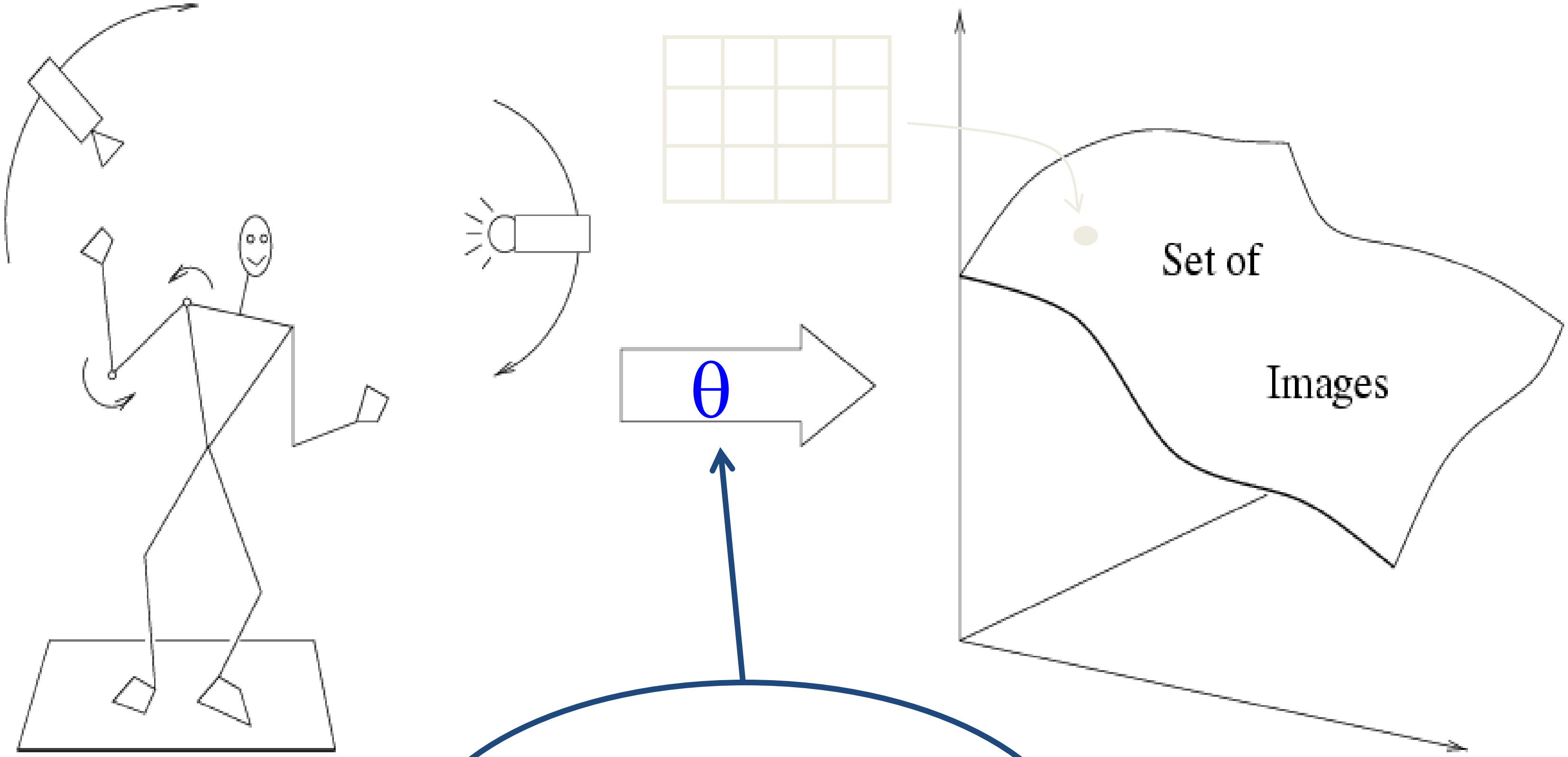
Within-class variations



Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era



Variability: Camera position
Illumination

Alignment

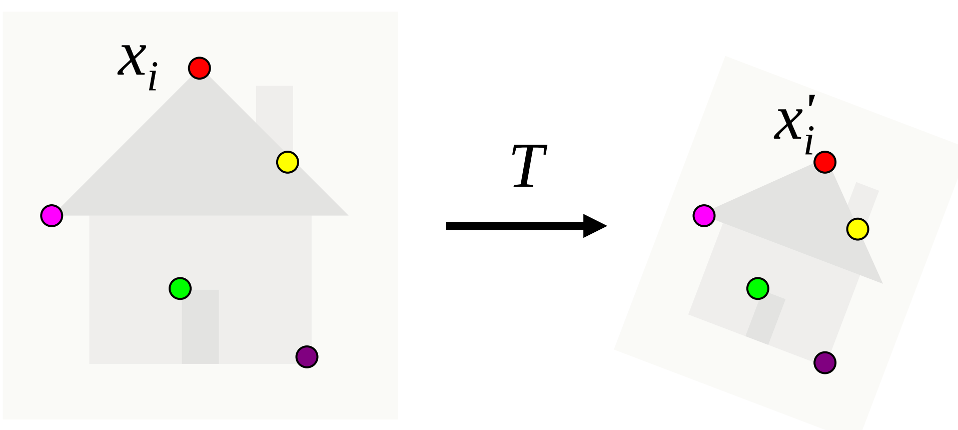
Shape: assumed known

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

Svetlana Lazebnik

Recall: Alignment

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

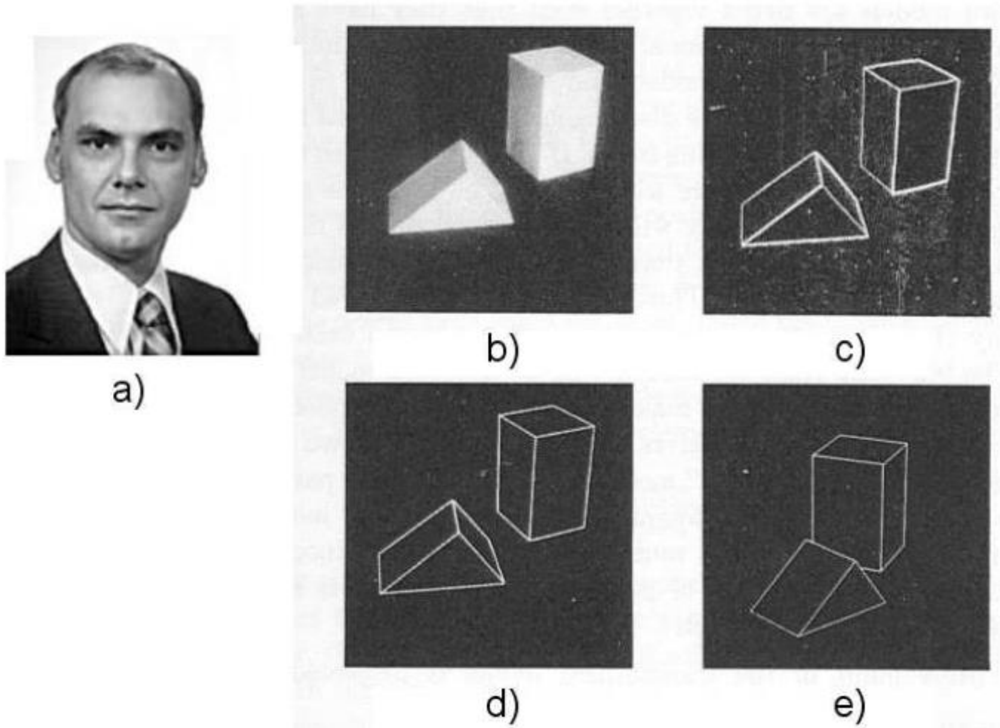


Find transformation T that minimizes

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

Svetlana Lazebnik

Recognition as an alignment problem: Block world

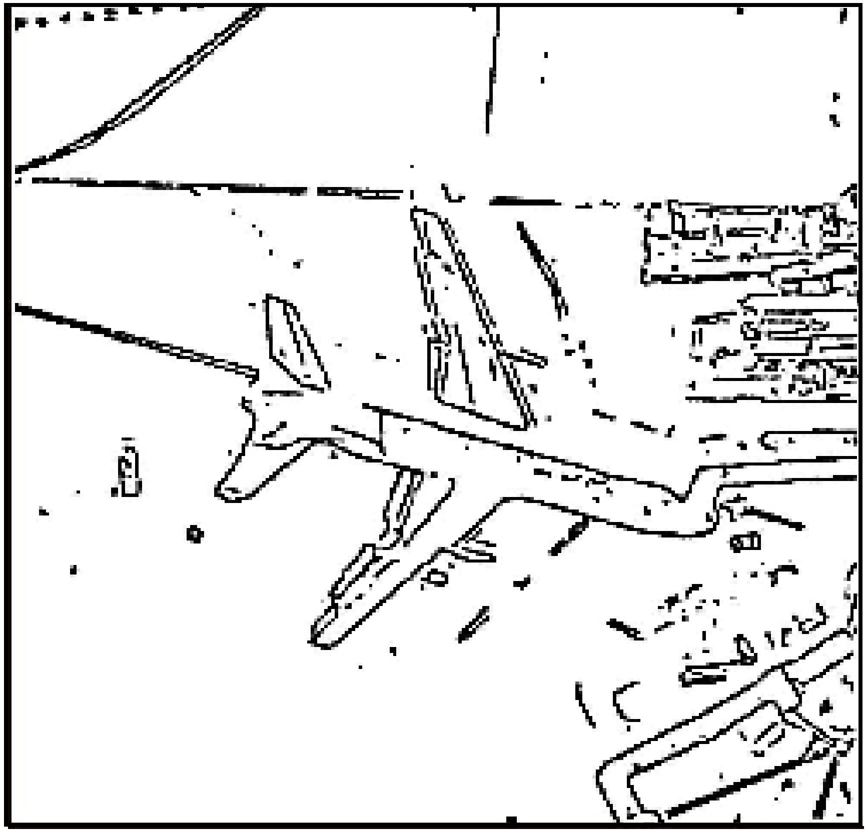


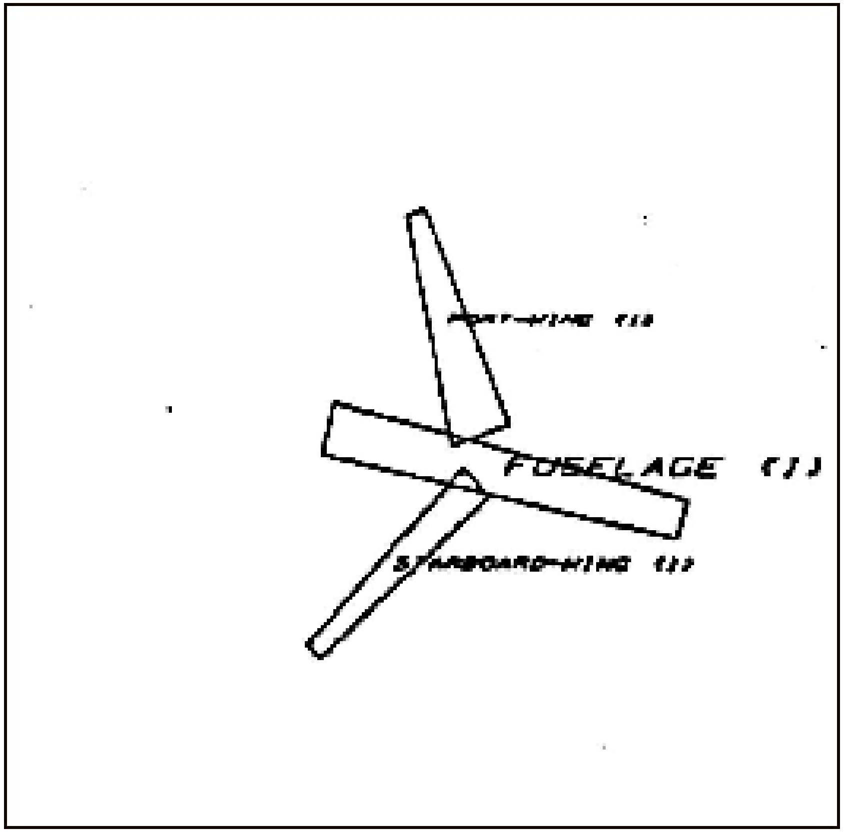
L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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

Representing and recognizing object categories is harder...





ACRONYM (Brooks and Binford, 1981)

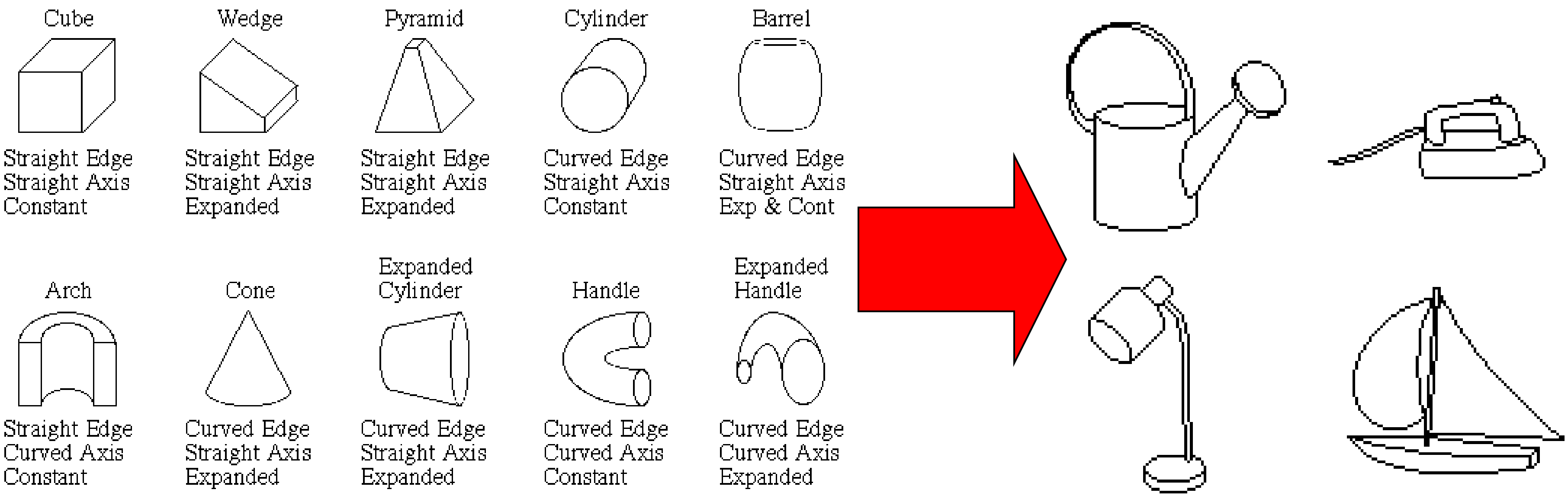
Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

Recognition by components

Biederman (1987)

Primitives (geons)

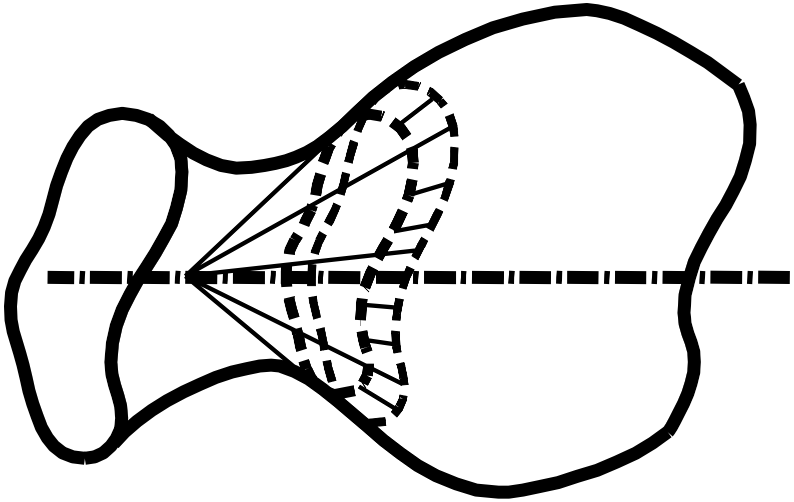
Objects



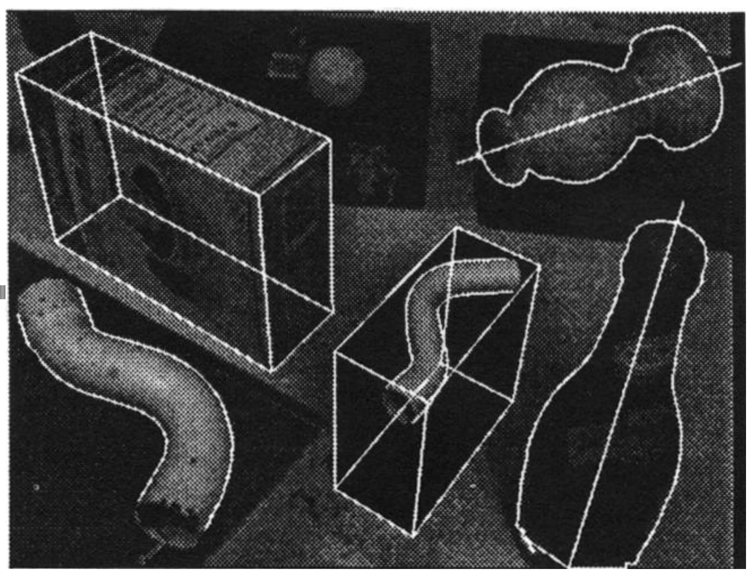
http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

Svetlana Lazebnik

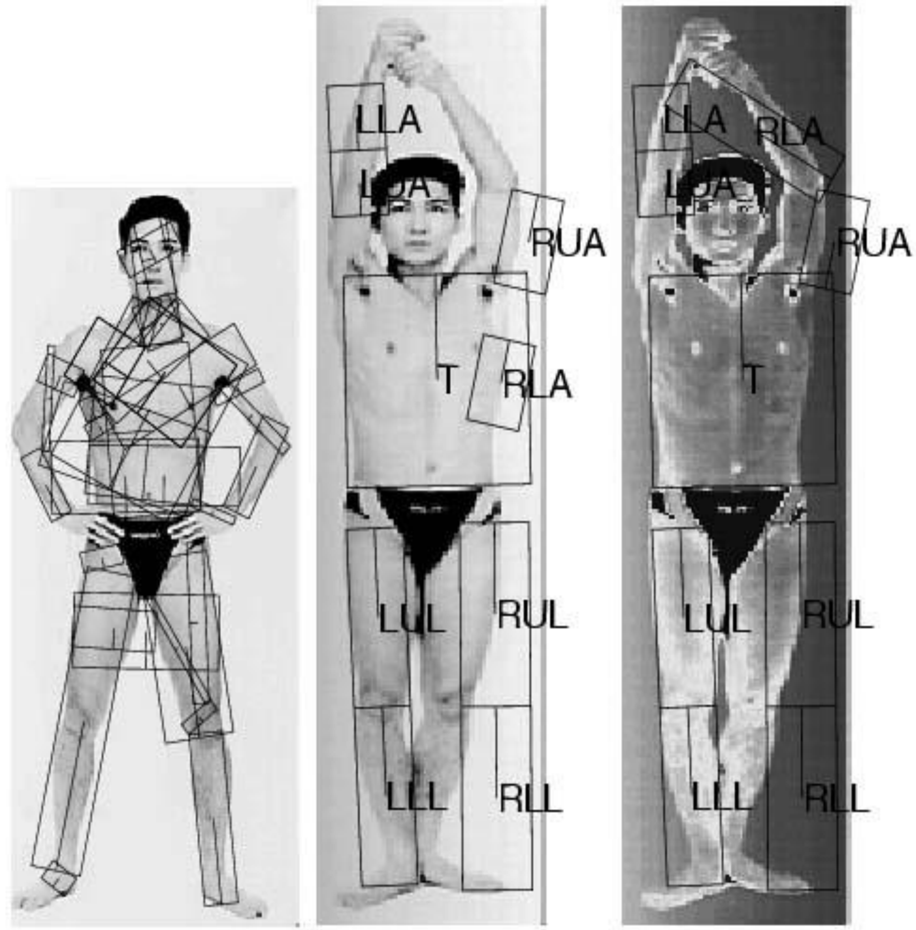
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



Zisserman et al. (1995)

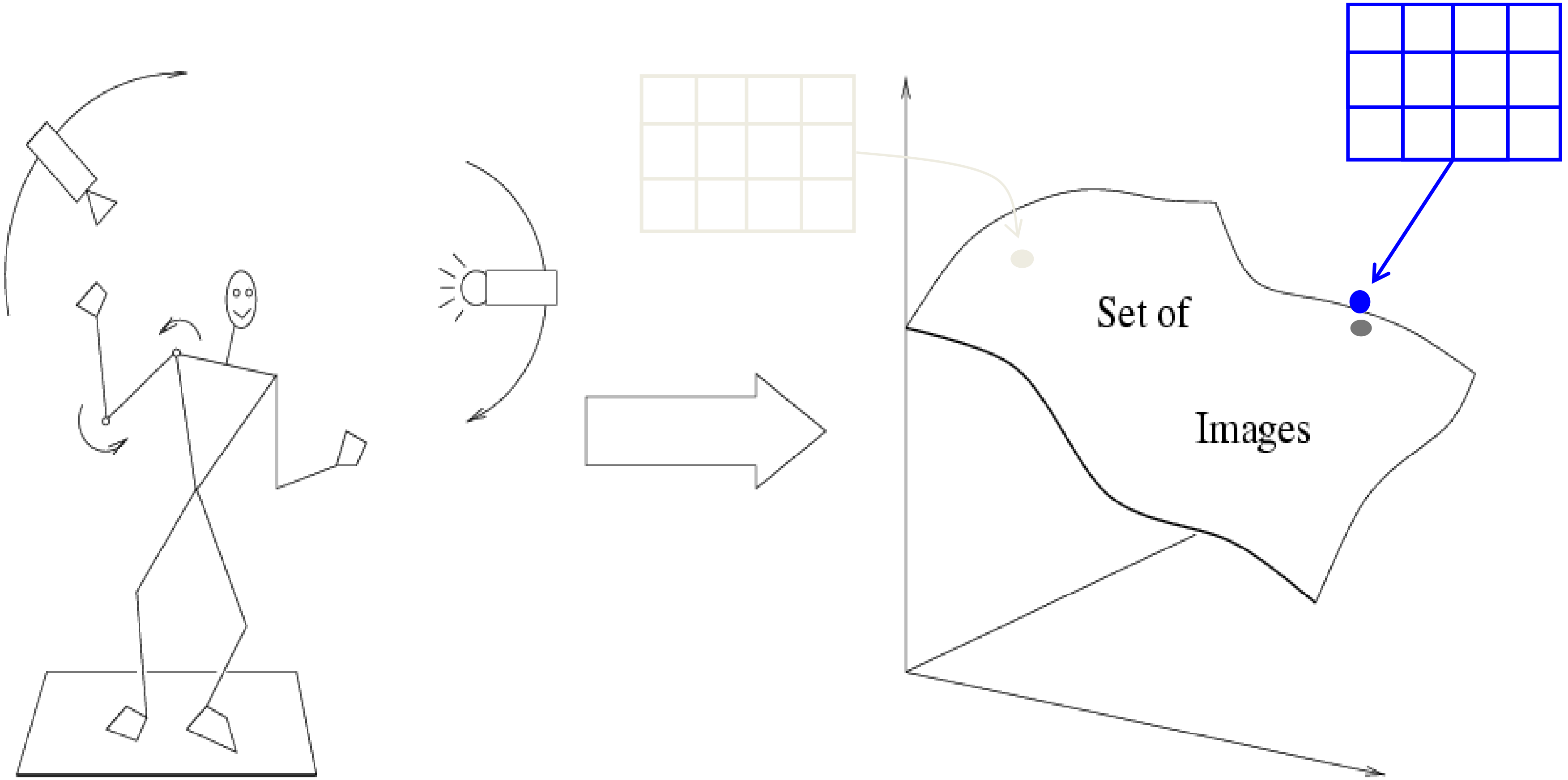


Forsyth (2000)

Svetlana Lazebnik

History of ideas in recognition

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



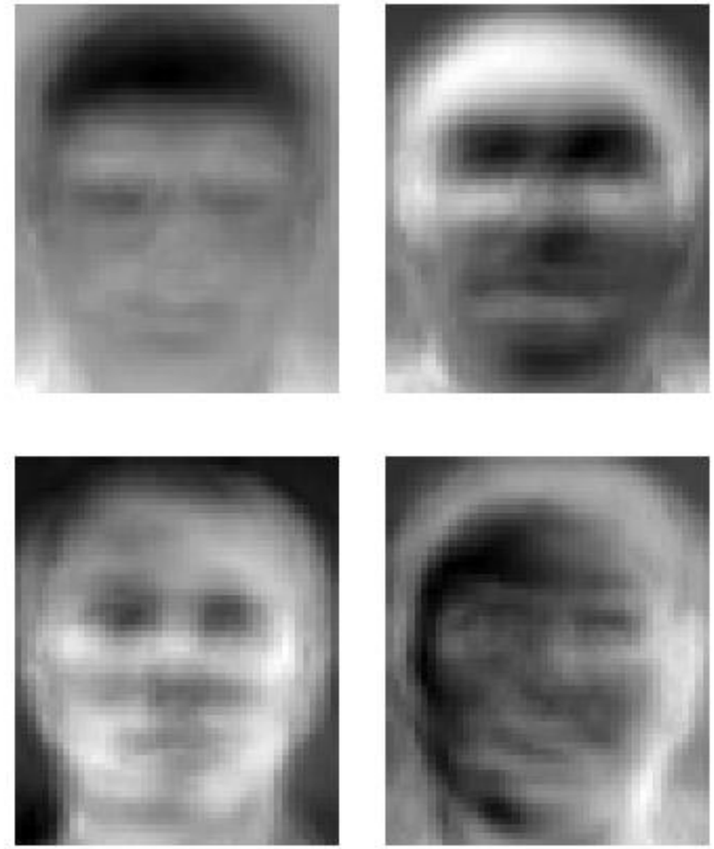
Empirical models of image variability

Appearance-based techniques

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

Svetlana Lazebnik

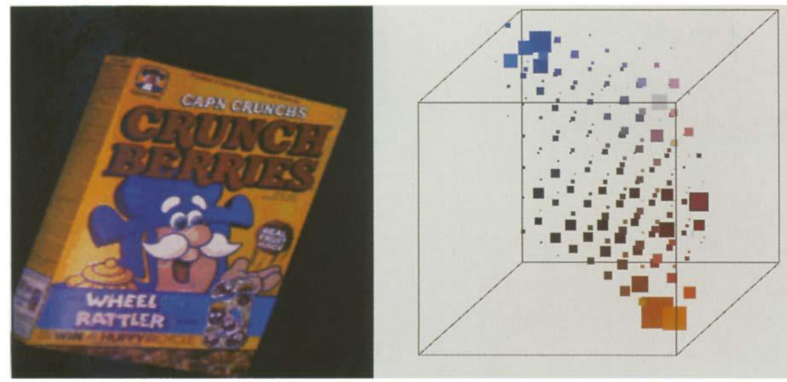
Eigenfaces (Turk & Pentland, 1991)

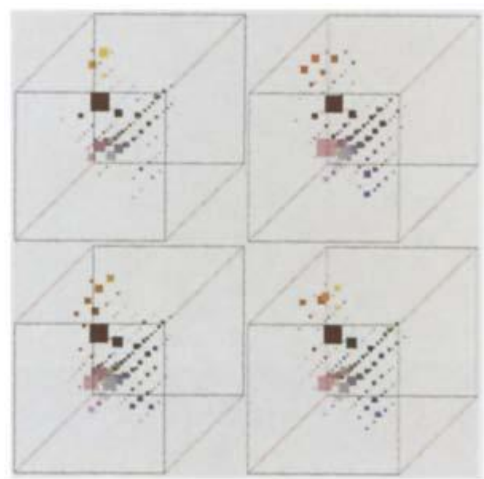


Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

Svetlana Lazebnik

Color Histograms

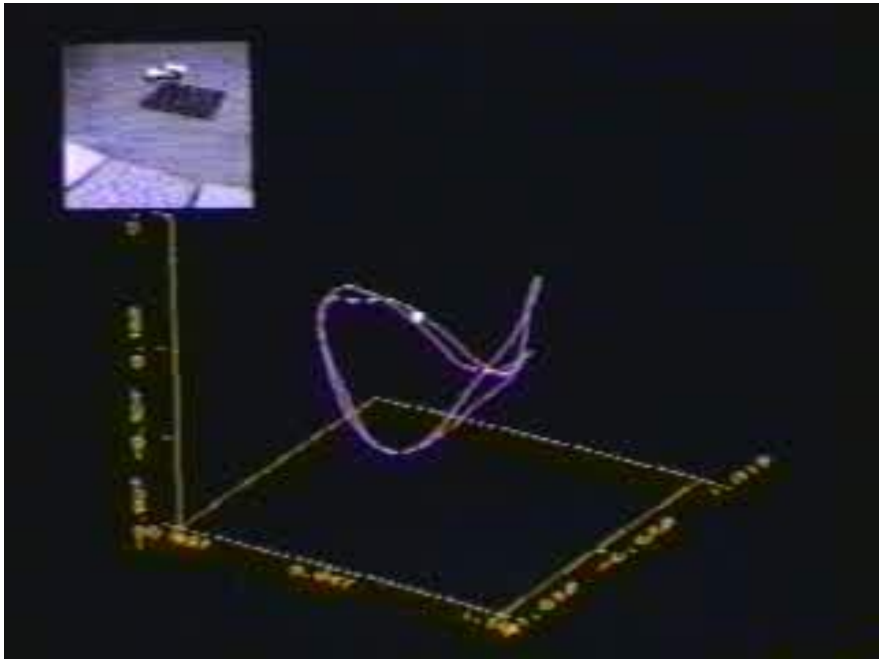




Swain and Ballard, Color Indexing, IJCV 1991.

Svetlana Lazebnik

Appearance manifolds



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

Limitations of global appearance models

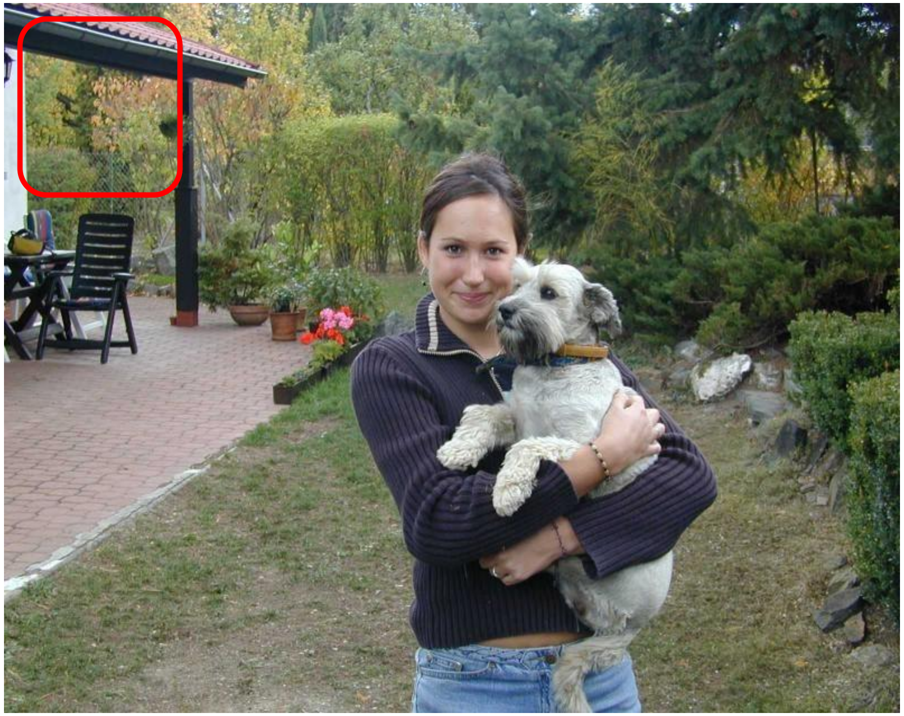
- Requires global registration of patterns
- Not robust to clutter, occlusion, geometric transformations

Svetlana Lazebnik

History of ideas in recognition

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

Sliding window approaches



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 object instance recognition

D. Lowe (1999, 2004)

Large-scale image search

Combining local features, indexing, and spatial constraints

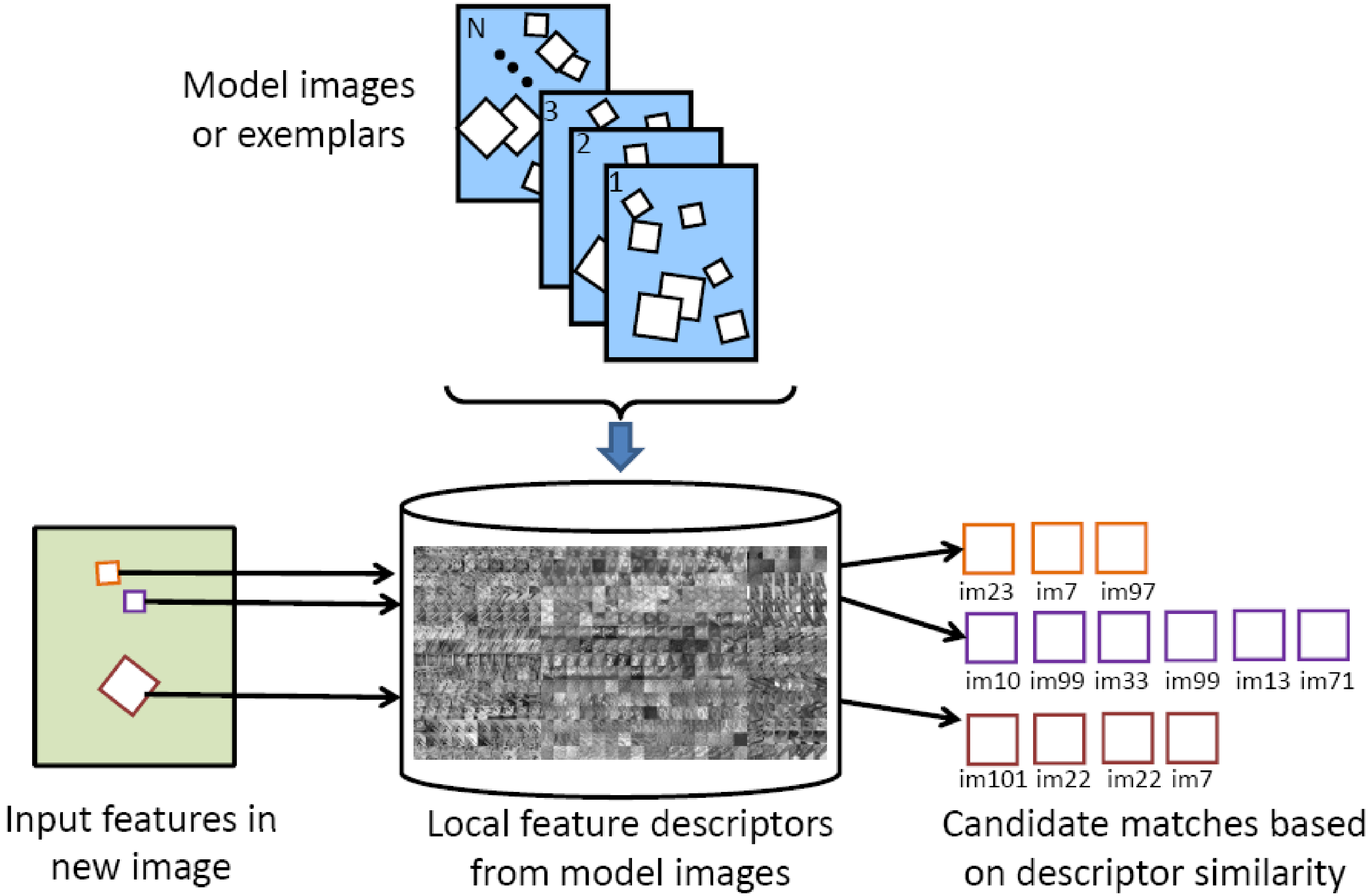
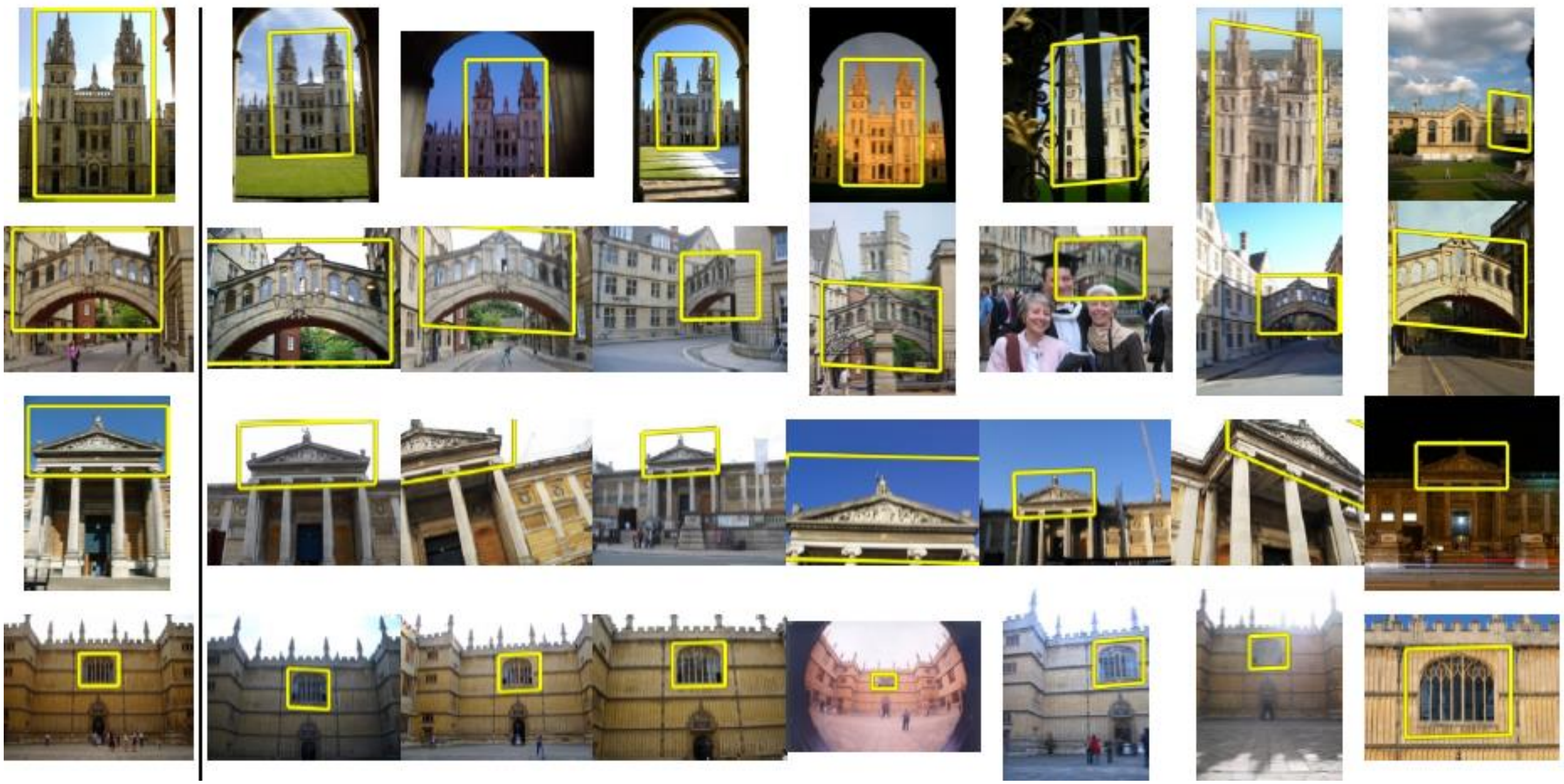


Image credit: K. Grauman and B. Leibe

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

Svetlana Lazebnik

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

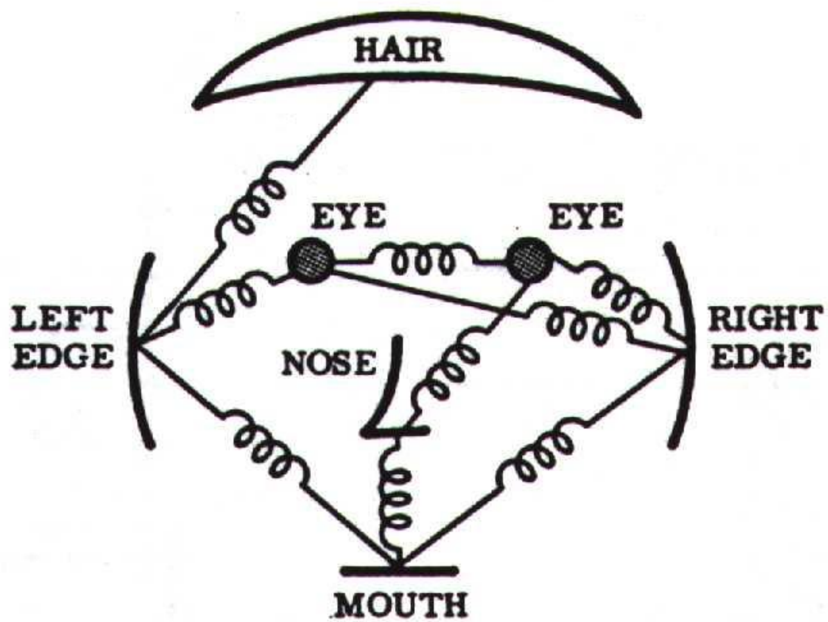


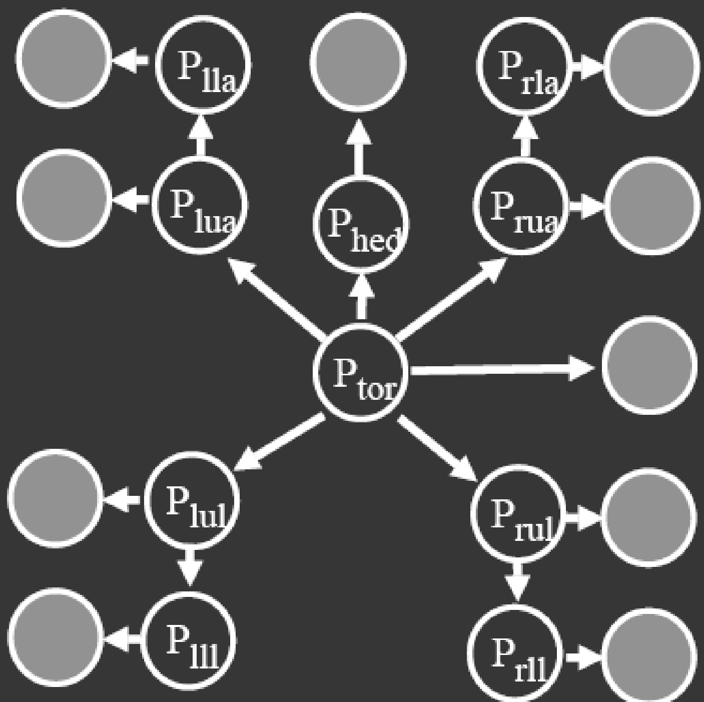
Figure from [Fischler & Elschlager 73]

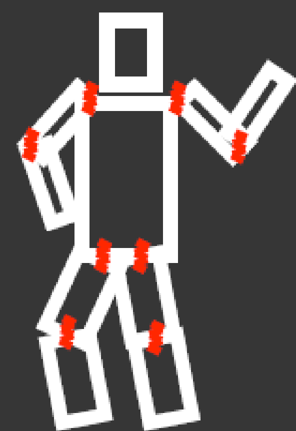
Constellation models

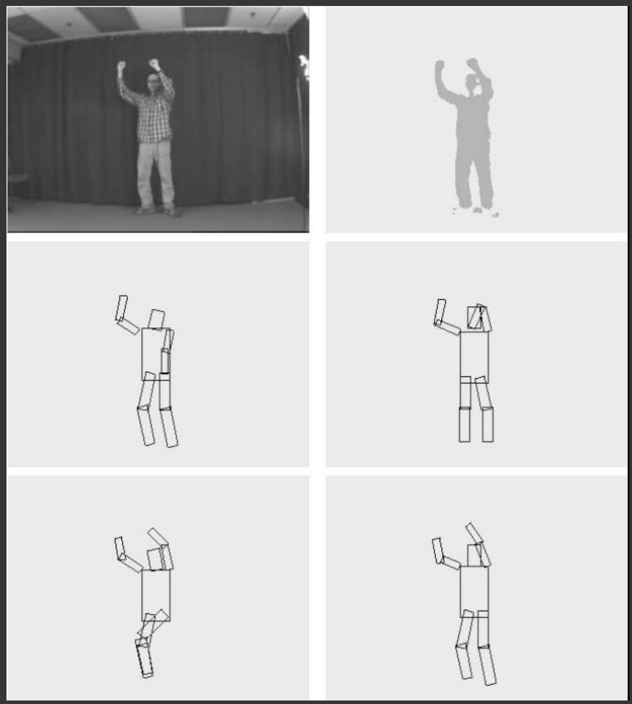
Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

Pictorial structure model

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





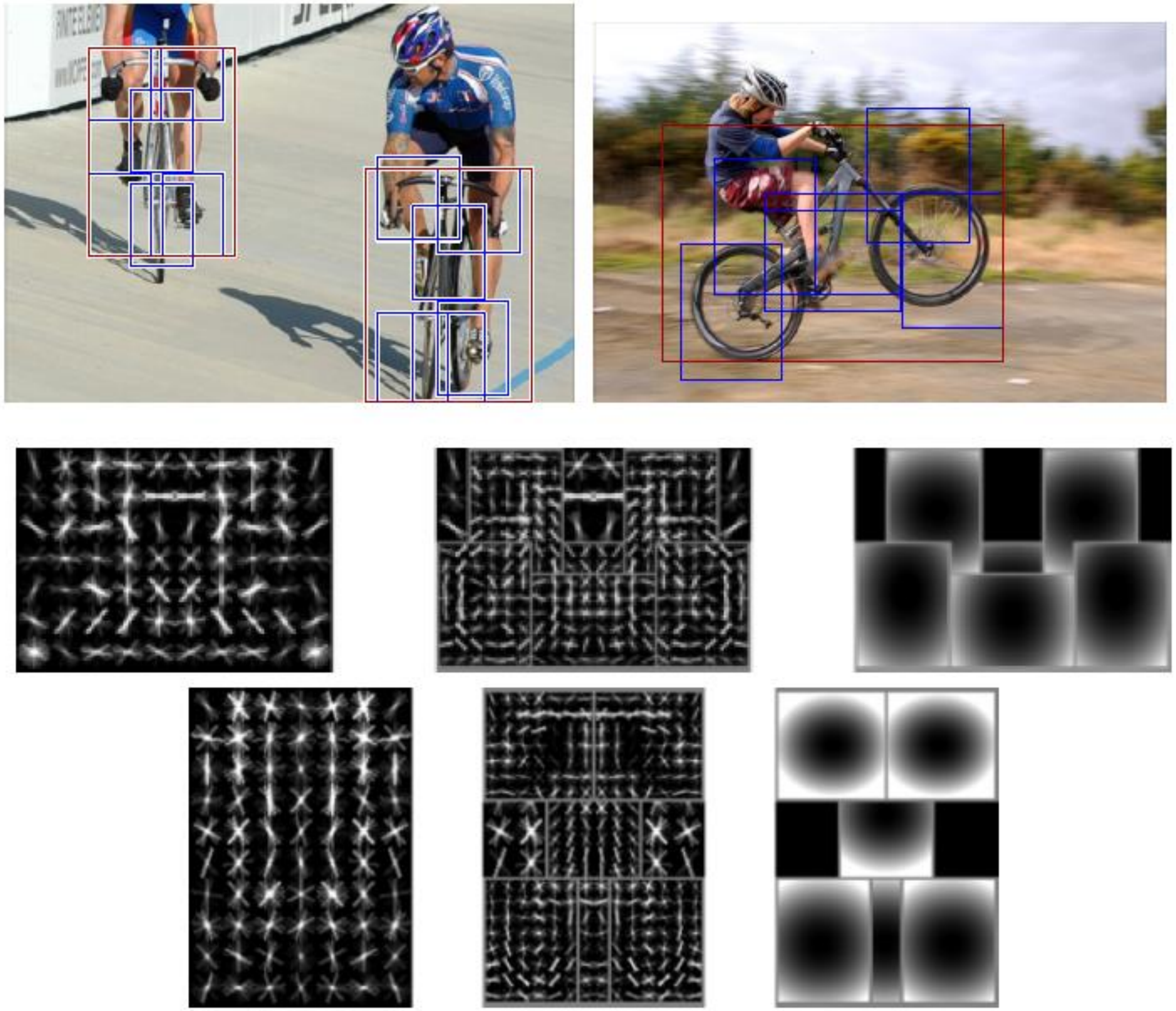


$$\Pr(P_{tor}, P_{arm}, \ldots | \mathrm{Im}) \propto \prod_{i,j} \Pr(P_i \mid P_j) \prod_i \Pr(\mathrm{Im}(P_i))$$

part geometry

part appearance

Discriminatively trained part-based models



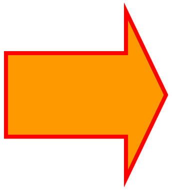
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," PAMI 2009

History of ideas in recognition

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

Svetlana Lazebnik

Bag-of-features models



Svetlana Lazebnik

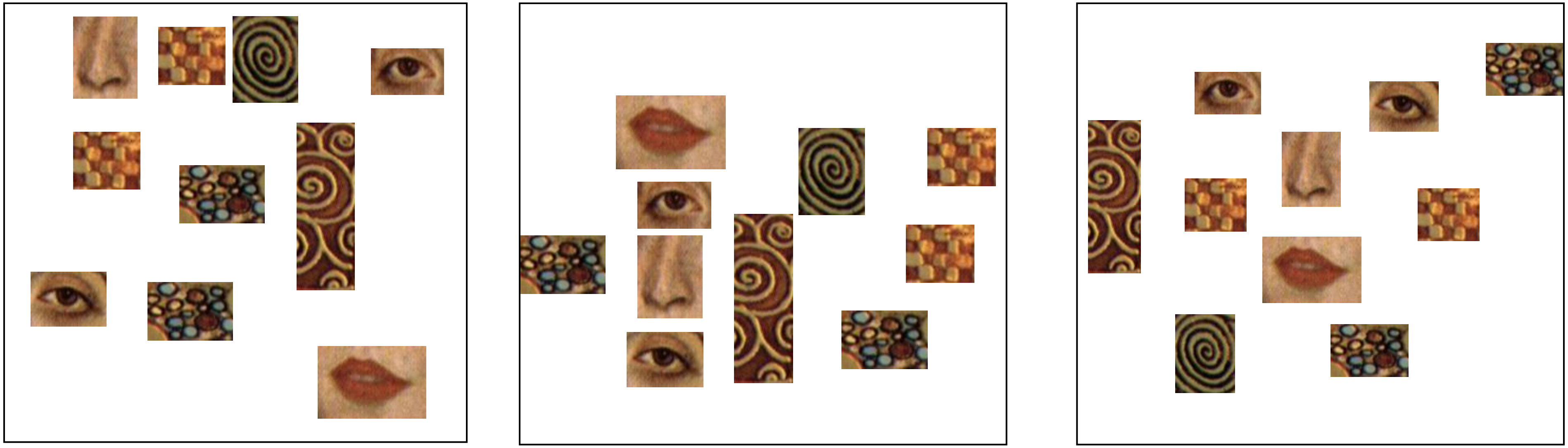
Bag-of-features models

Object → **Bag of ‘words’**

Svetlana Lazebnik

Objects as texture

- All of these are treated as being the same

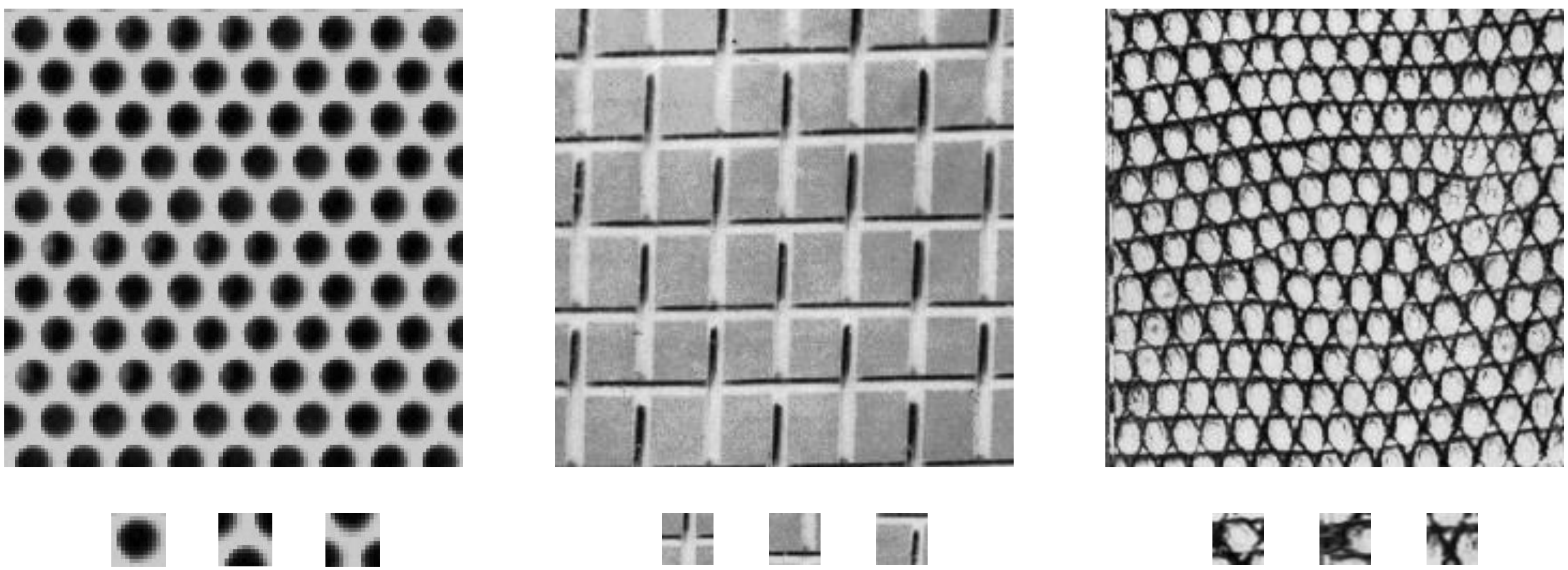


- No distinction between foreground and background: scene recognition?

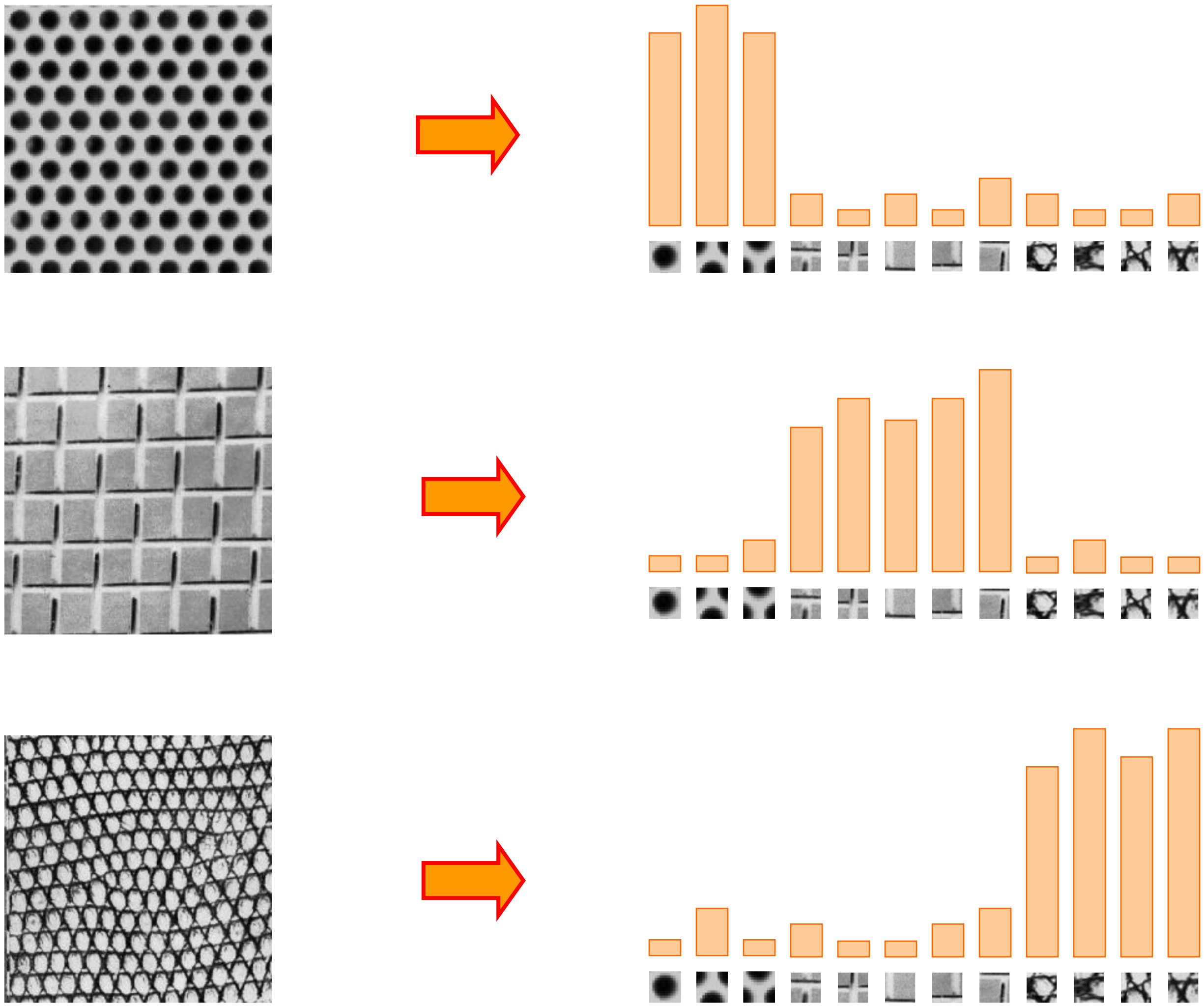
Svetlana Lazebnik

Origin 1: Texture recognition

- Texture is characterized by the repetition of basic elements or *textons*
- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



Origin 1: Texture recognition



Origin 2: Bag-of-words models

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

Origin 2: Bag-of-words models

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

Origin 2: Bag-of-words models

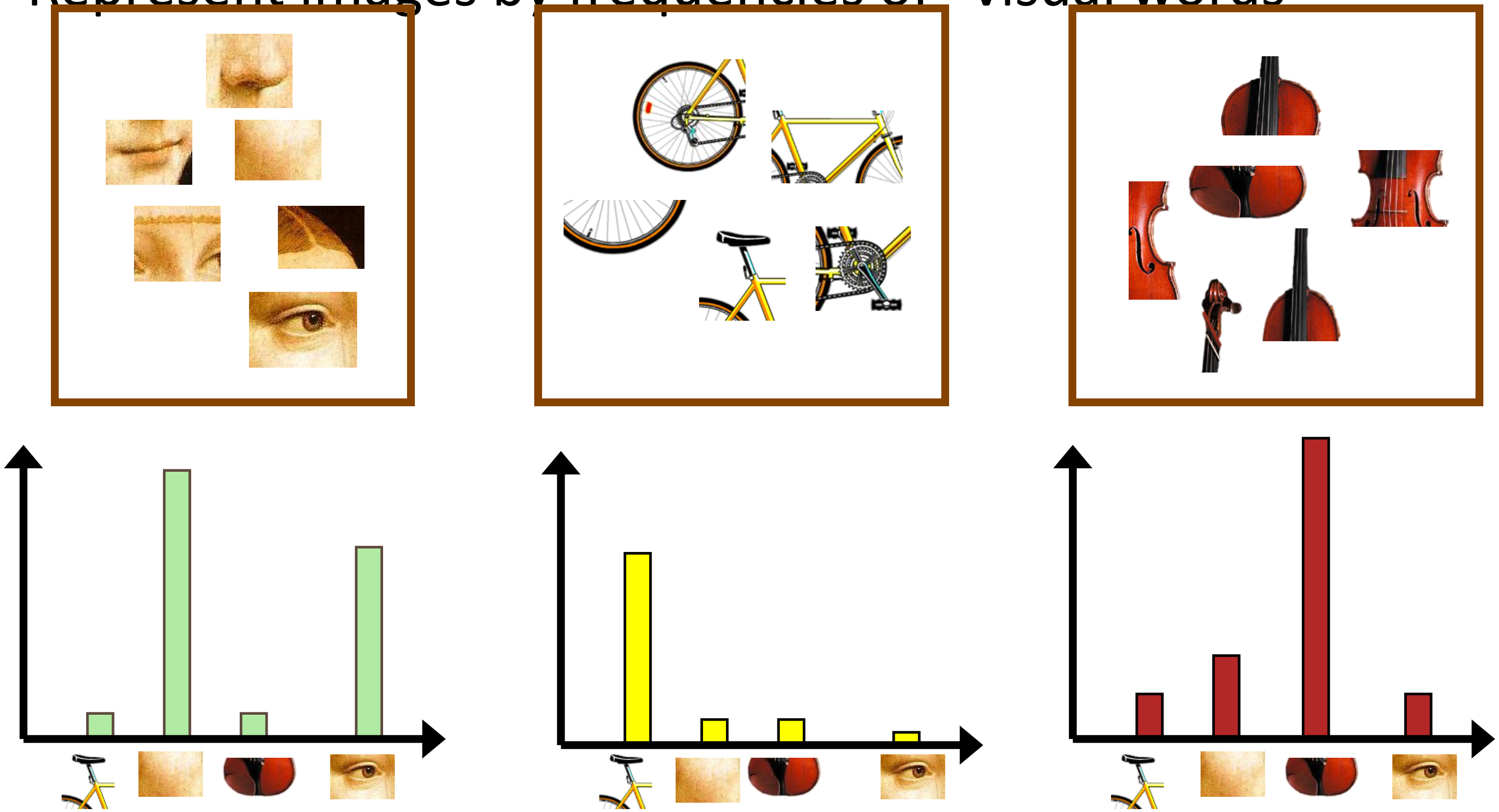
- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

Origin 2: Bag-of-words models

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

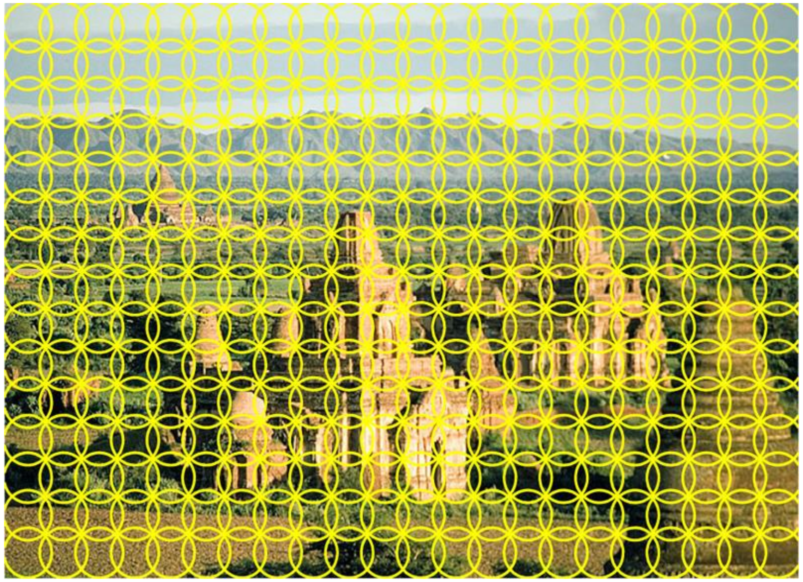
Bag-of-features steps

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

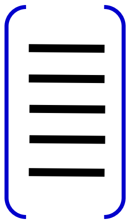


1. Feature extraction

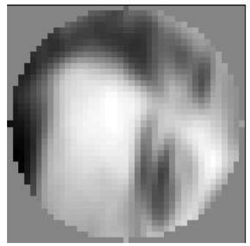
- Regular grid or interest regions



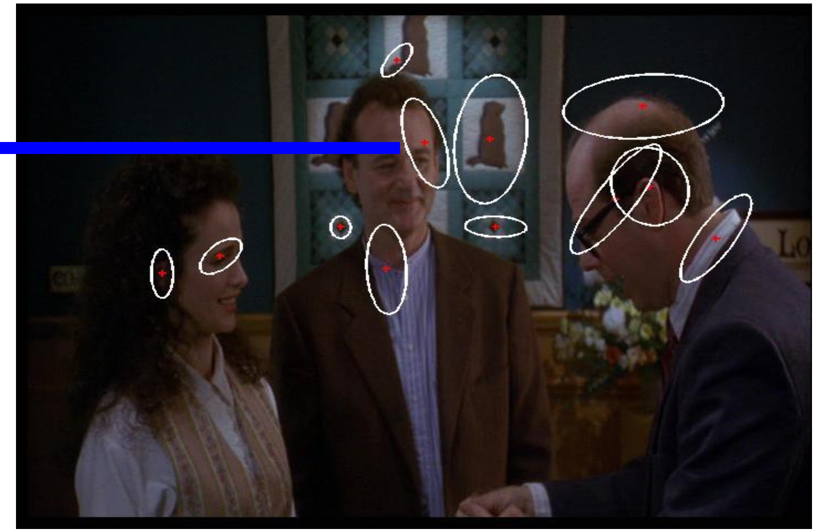
1. Feature extraction



Compute descriptor



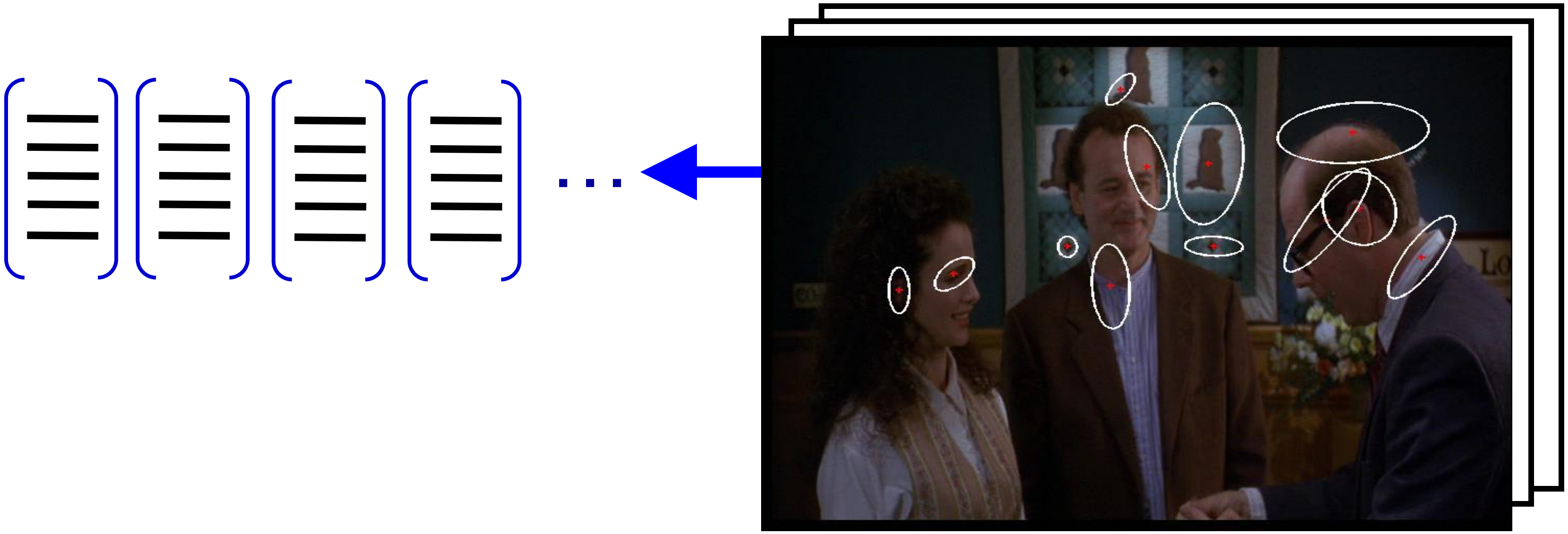
Normalize patch



Detect patches

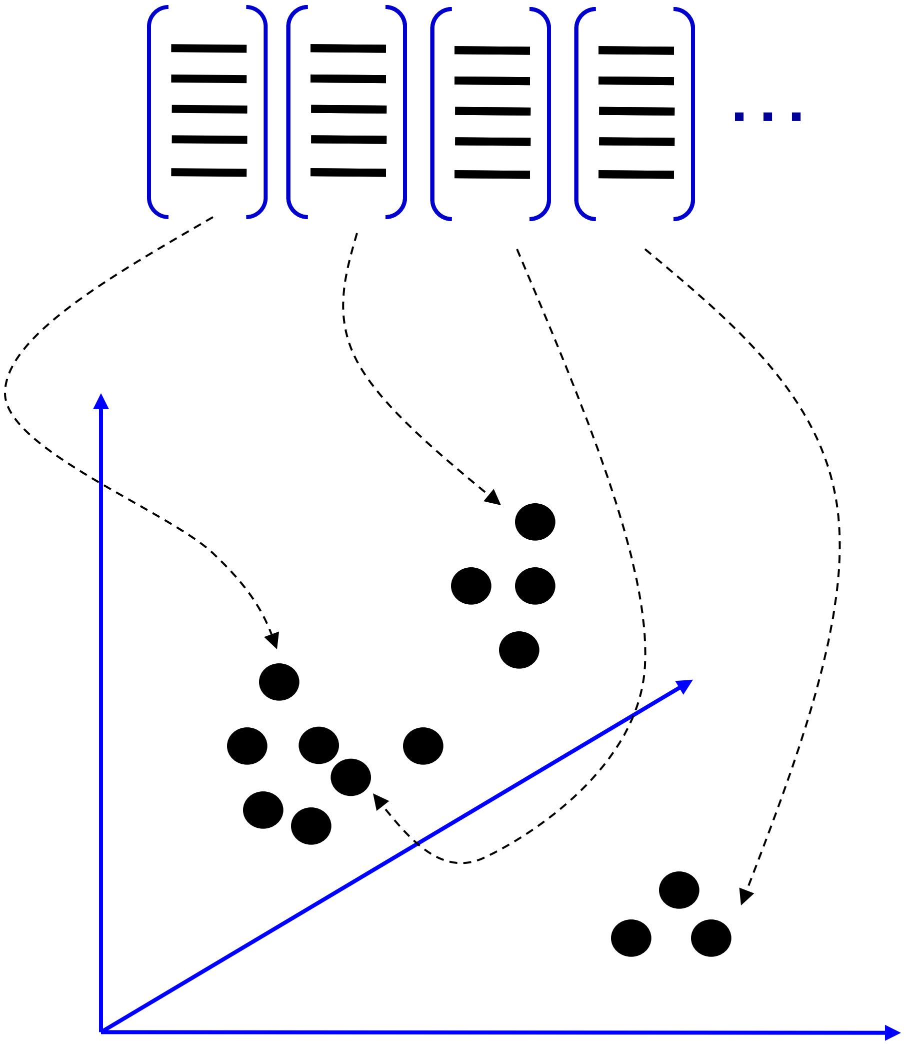
Slide credit: Josef Sivic

1. Feature extraction



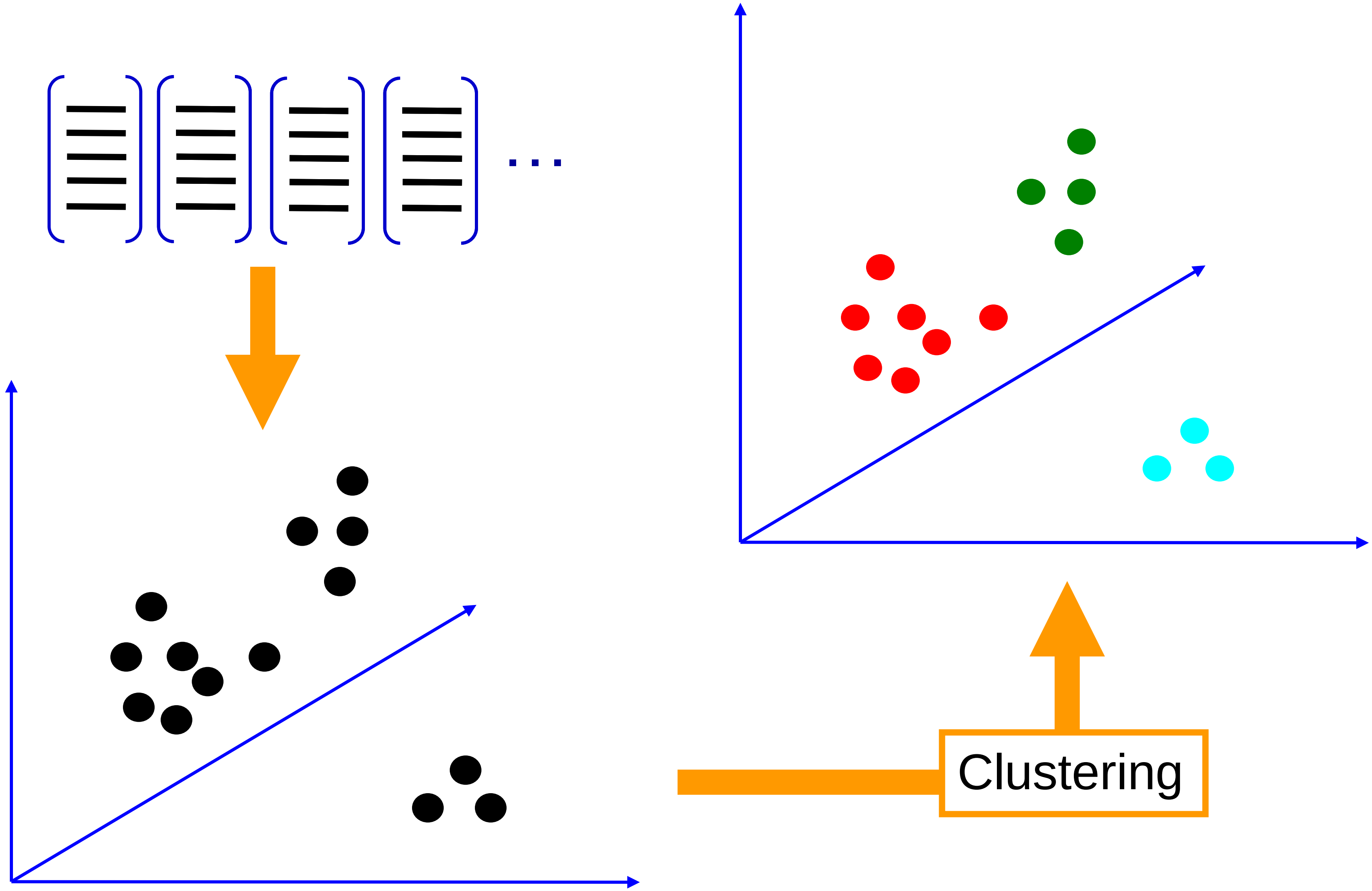
Slide credit: Josef Sivic

2. Learning the visual vocabulary



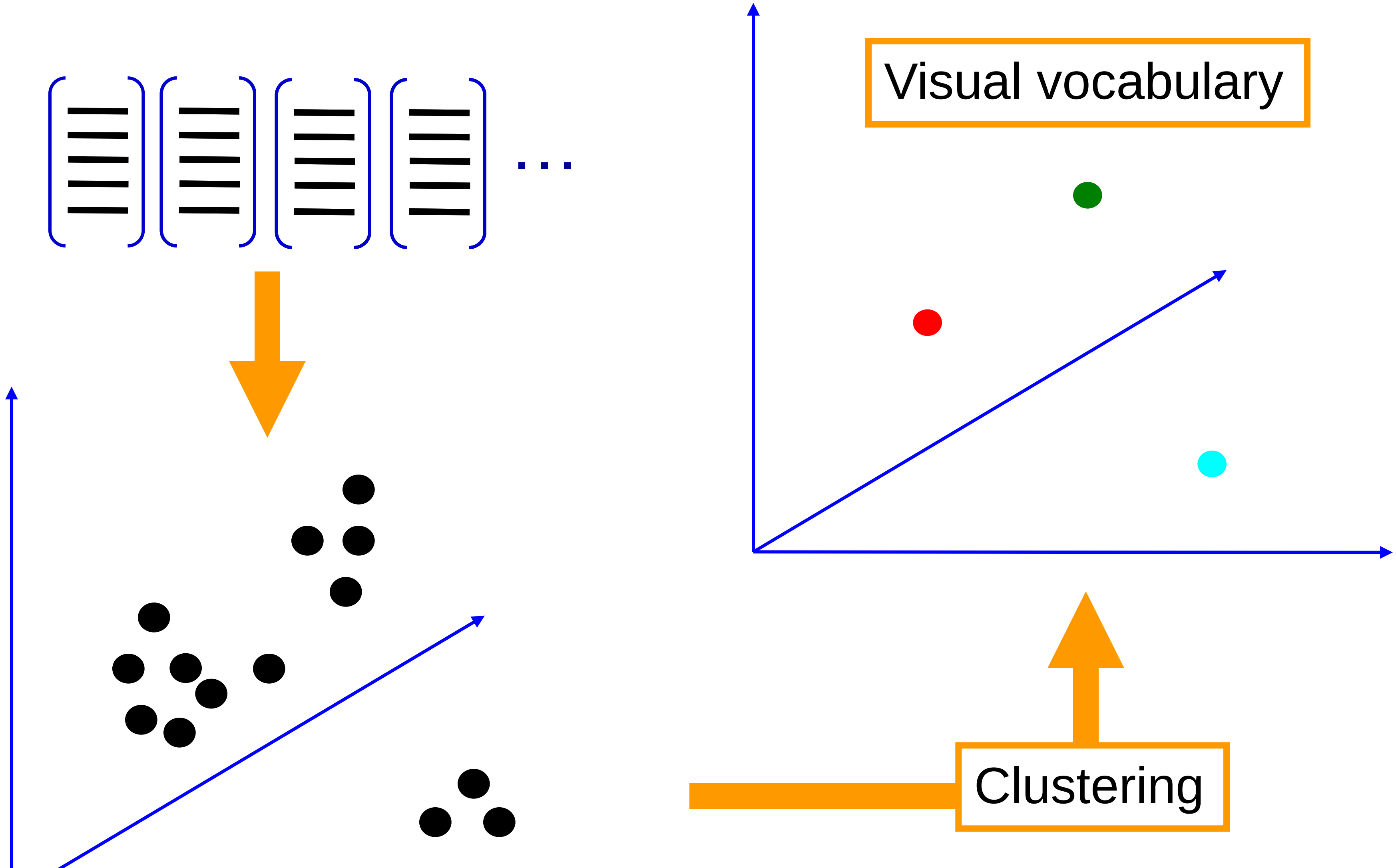
Slide credit: Josef Sivic

2. Learning the visual vocabulary



Slide credit: Josef Sivic

2. Learning the visual vocabulary



Slide credit: Josef Sivic

K-means clustering

- Want to minimize sum of squared Euclidean distances between points x_i and their nearest cluster centers m_k

$$D(X, M) = \sum_{\text{cluster } k} \sum_{\substack{\text{point } i \text{ in} \\ \text{cluster } k}} (x_i - m_k)^2$$

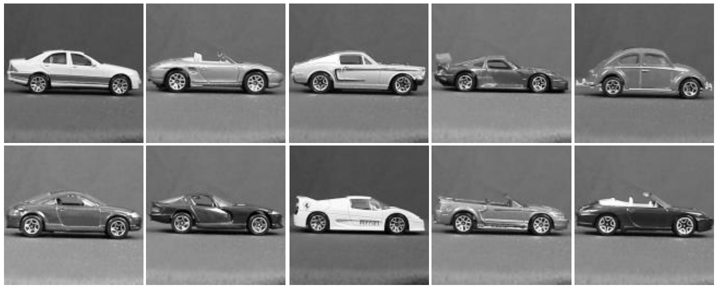
- Algorithm:
- Randomly initialize K cluster centers
- Iterate until convergence:
  - Assign each data point to the nearest center
  - Recompute each cluster center as the mean of all points assigned to it

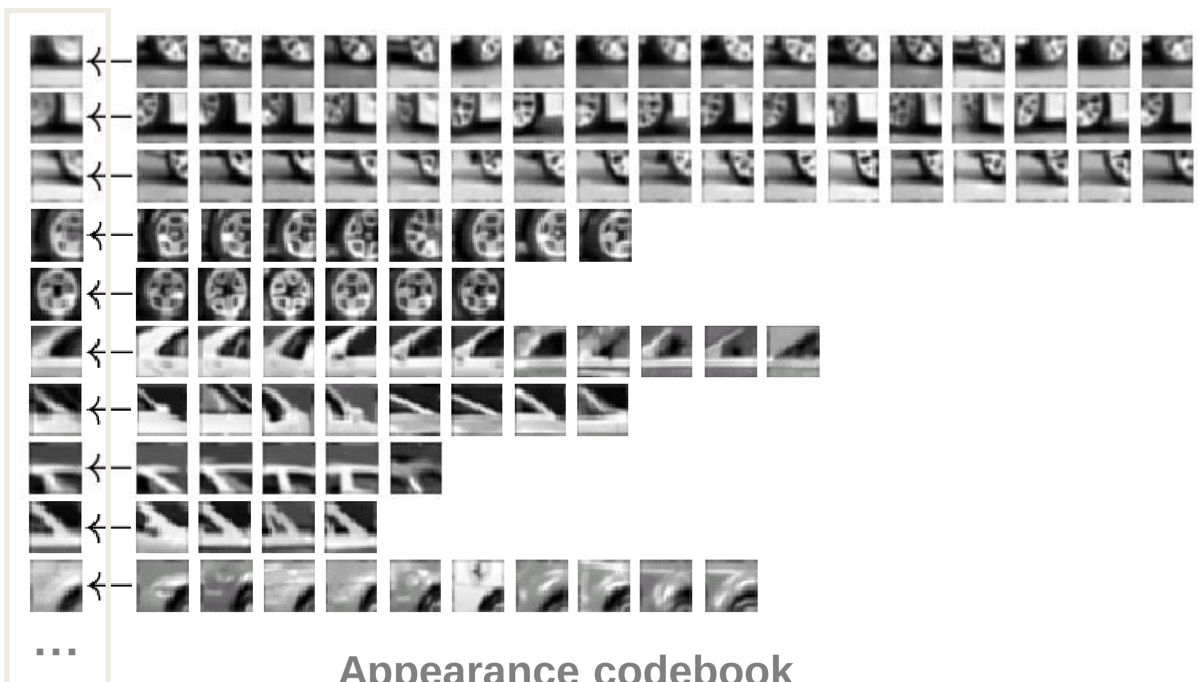
Clustering and vector quantization

- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process
  - Each cluster center produced by k-means becomes a codevector
  - Codebook can be learned on separate training set
  - Provided 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

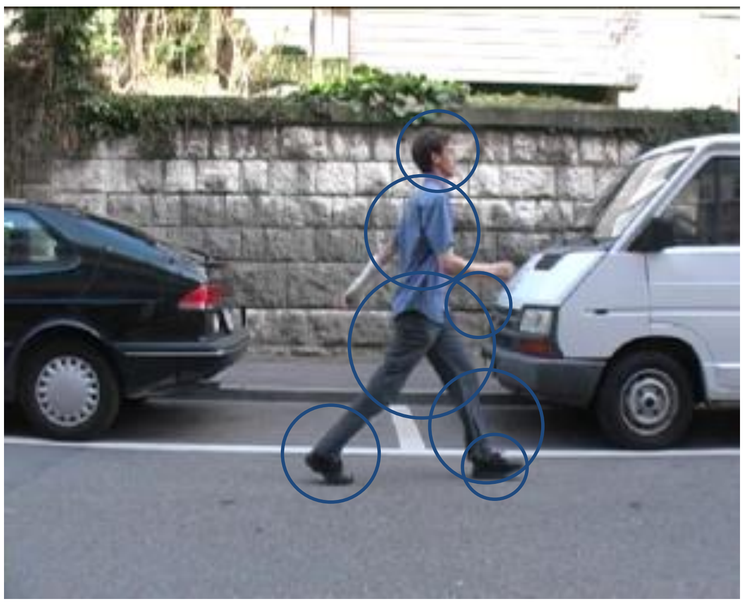
Example codebook





Source: B. Leibe

Another codebook

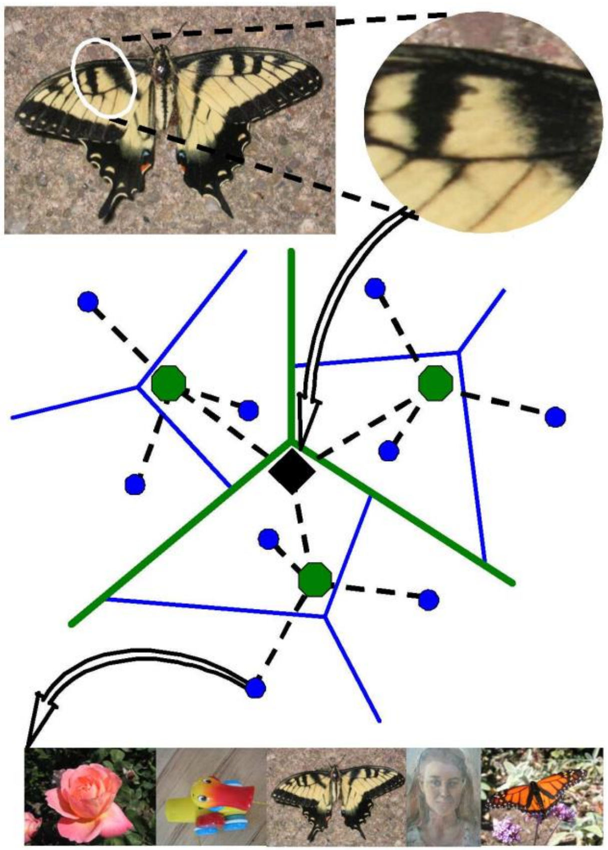


Appearance codebook

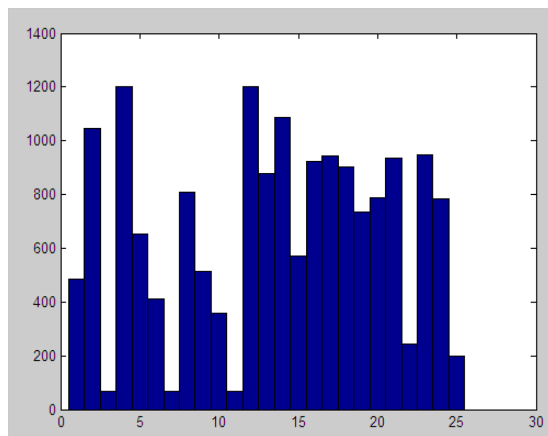
Source: B. Leibe

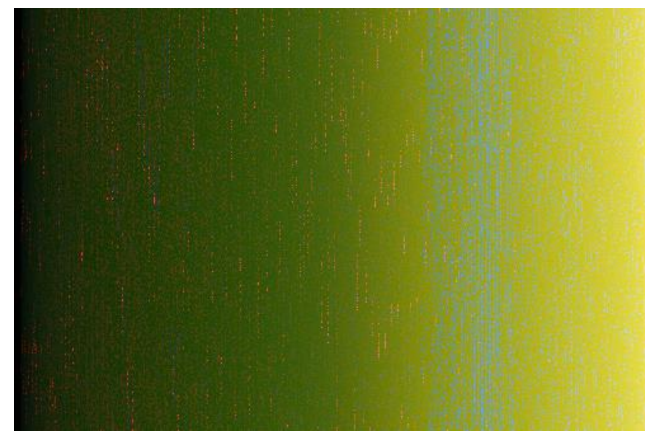
Visual vocabularies: Issues

- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large: quantization artifacts, overfitting
- Computational efficiency
  - Vocabulary trees
    (Nister & Stewenius, 2006)



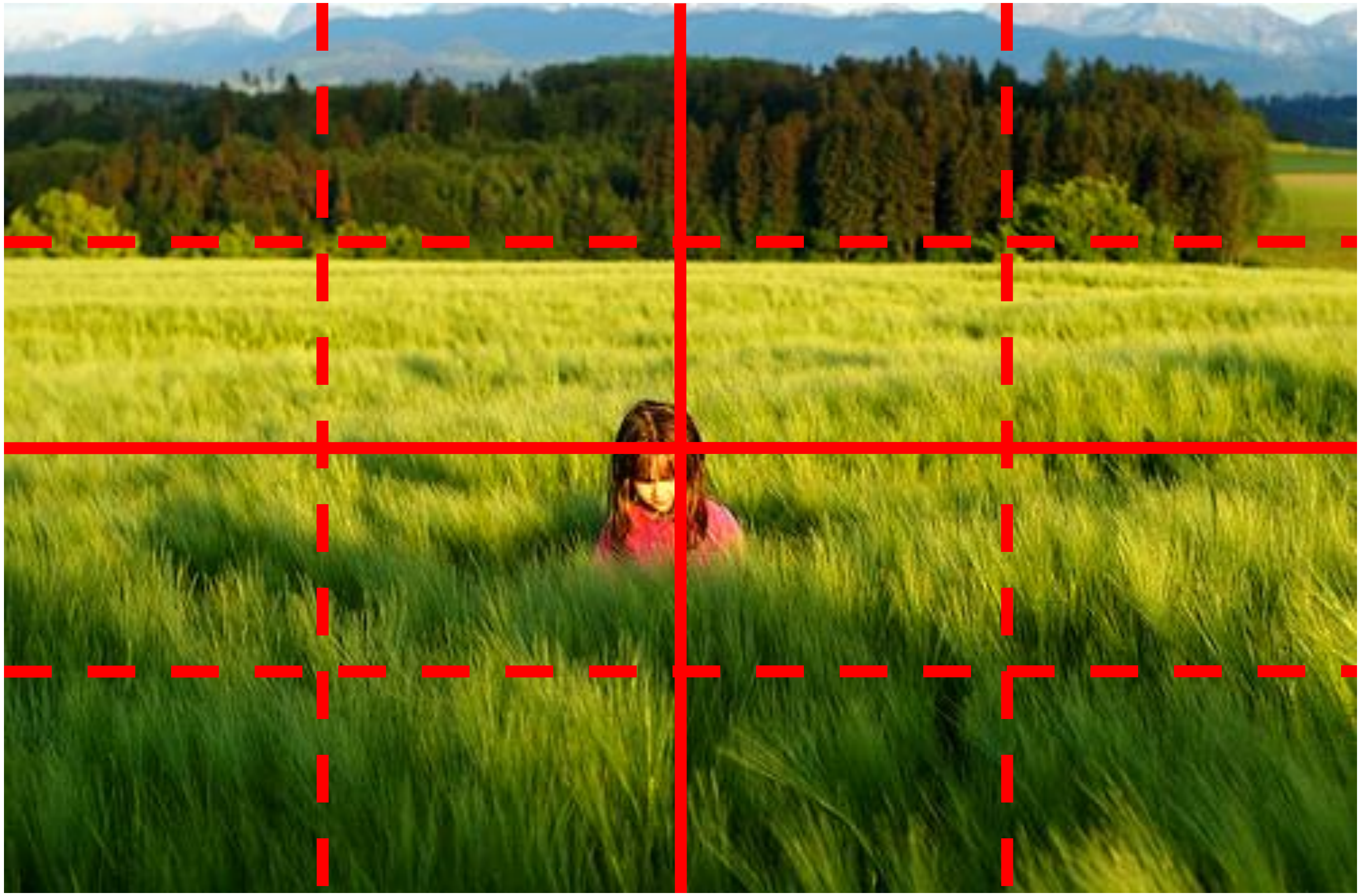
But what about layout?





All of these images have the same color histogram

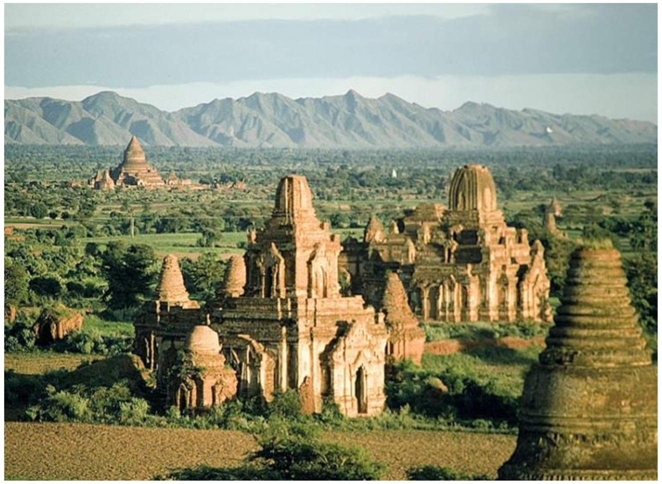
Spatial pyramid

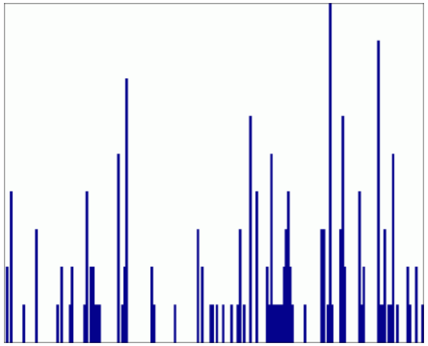


Compute histogram in each spatial bin

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

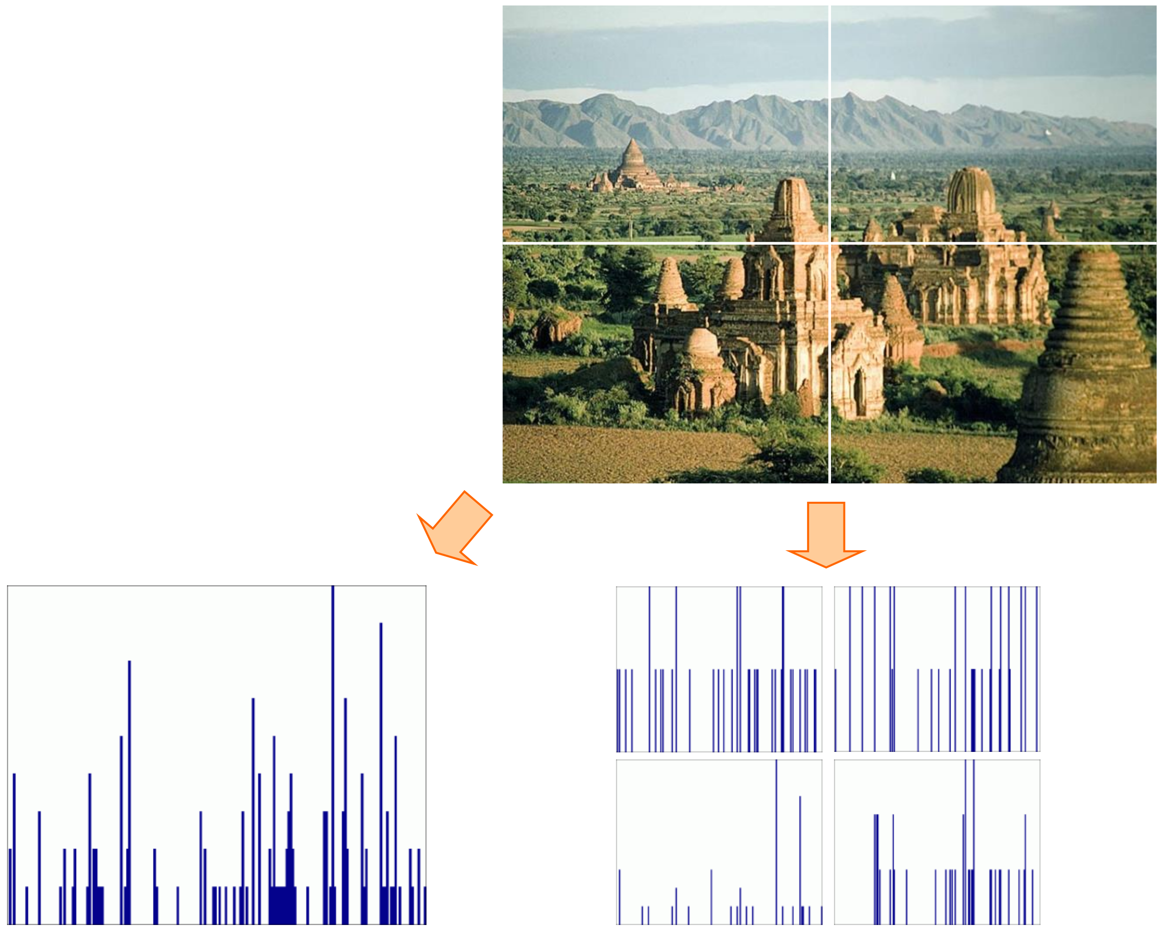




Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

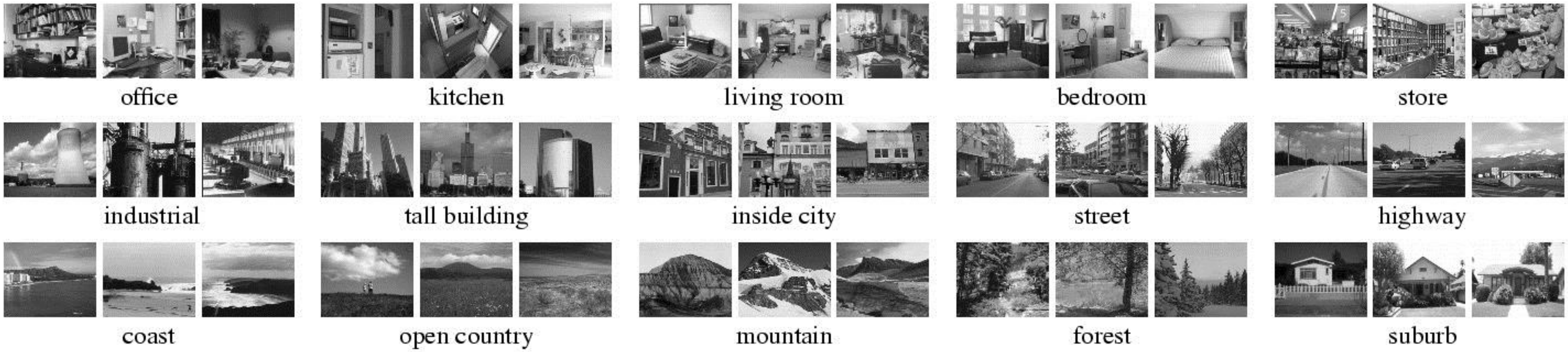


Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

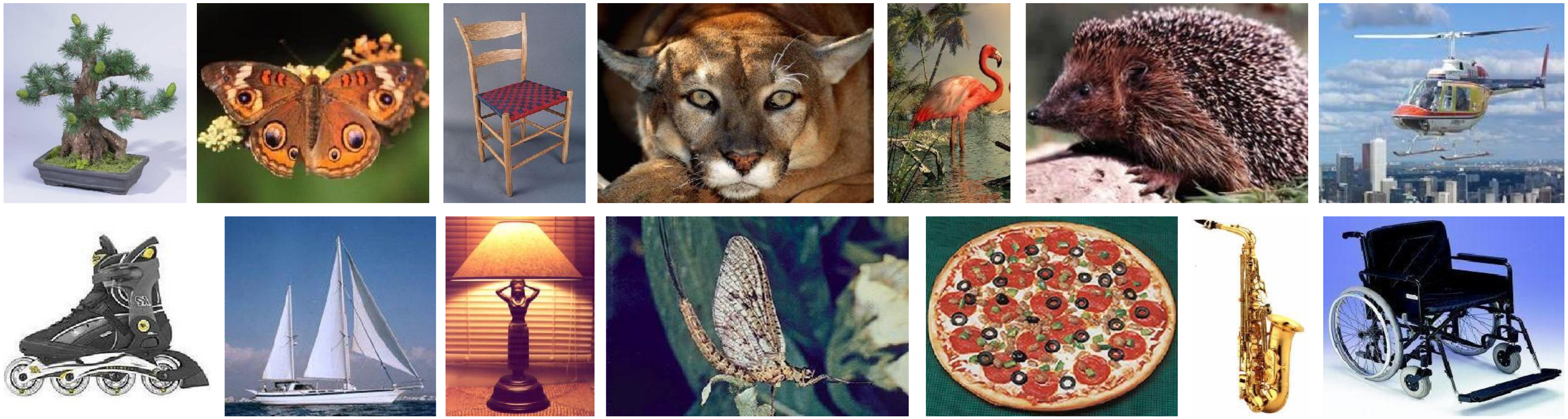
Lazebnik, Schmid & Ponce (CVPR 2006)



Multi-class classification results

	Weak features (vocabulary size: 16)		Strong features (vocabulary size: 200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6	
1 (2×2)	53.6 ±0.3	56.2 ±0.6	77.9 ±0.6	79.0 ±0.5
2 (4×4)	61.7 ±0.6	64.7 ±0.7	79.4 ±0.3	**81.1** ±0.3
3 (8×8)	63.3 ±0.8	**66.8** ±0.6	77.2 ±0.4	80.7 ±0.3

Caltech101 dataset

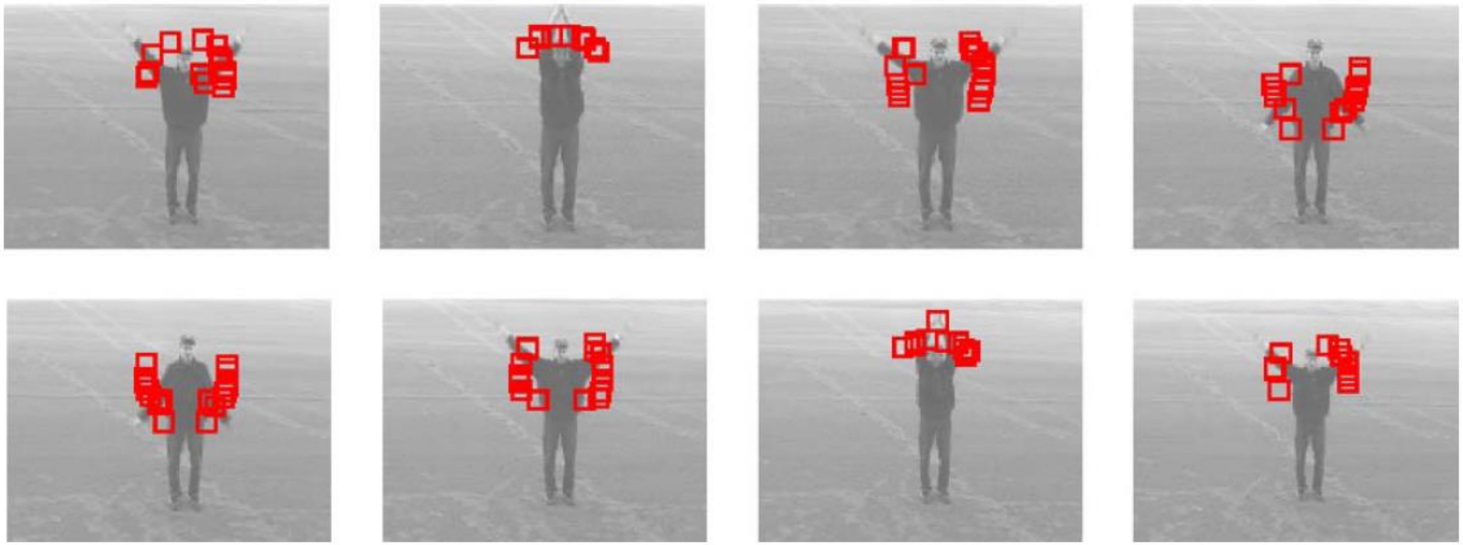


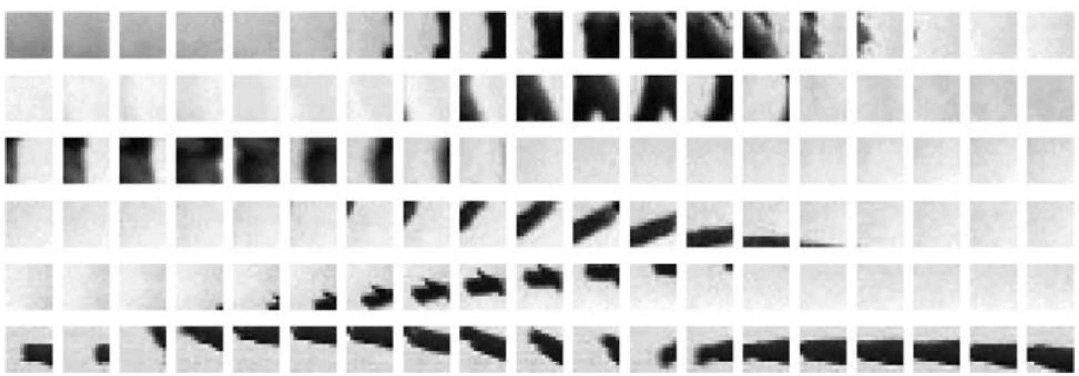
Multi-class classification results (30 training images per class)

	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

Bags of features for action recognition

Space-time interest points





Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei, **Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words**, IJCV 2008.

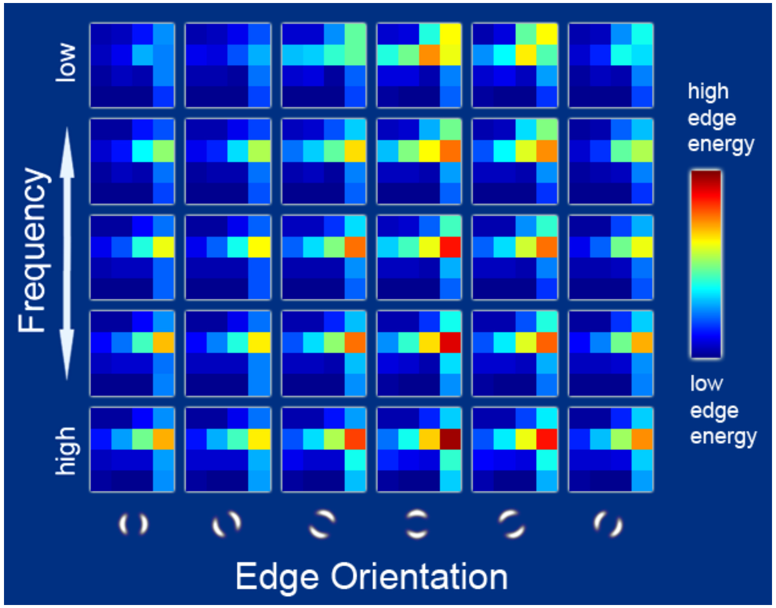
History of ideas in recognition

- 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
- Present trends: combination of local and global methods, data-driven methods, context

Svetlana Lazebnik

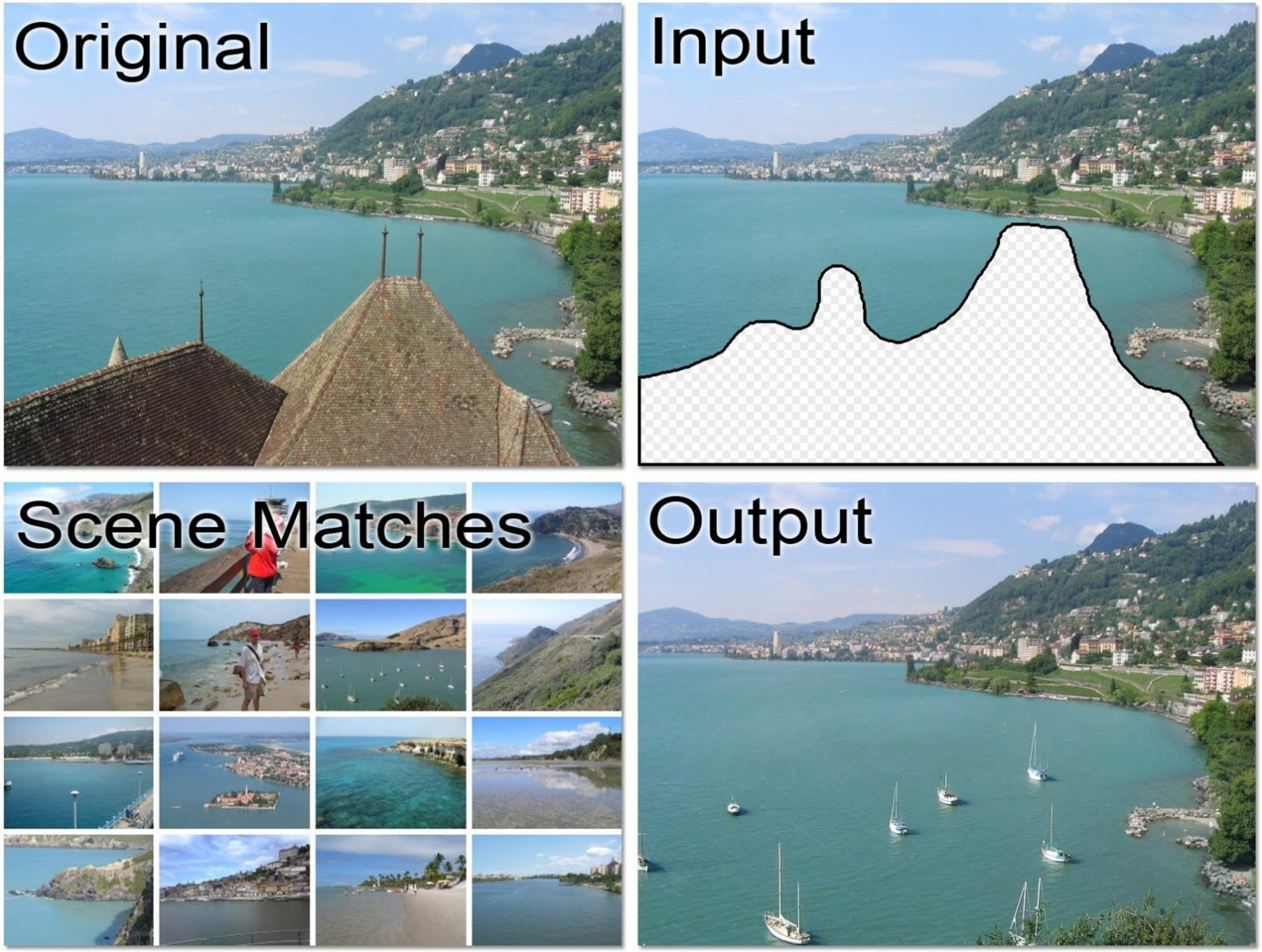
Global scene descriptors

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



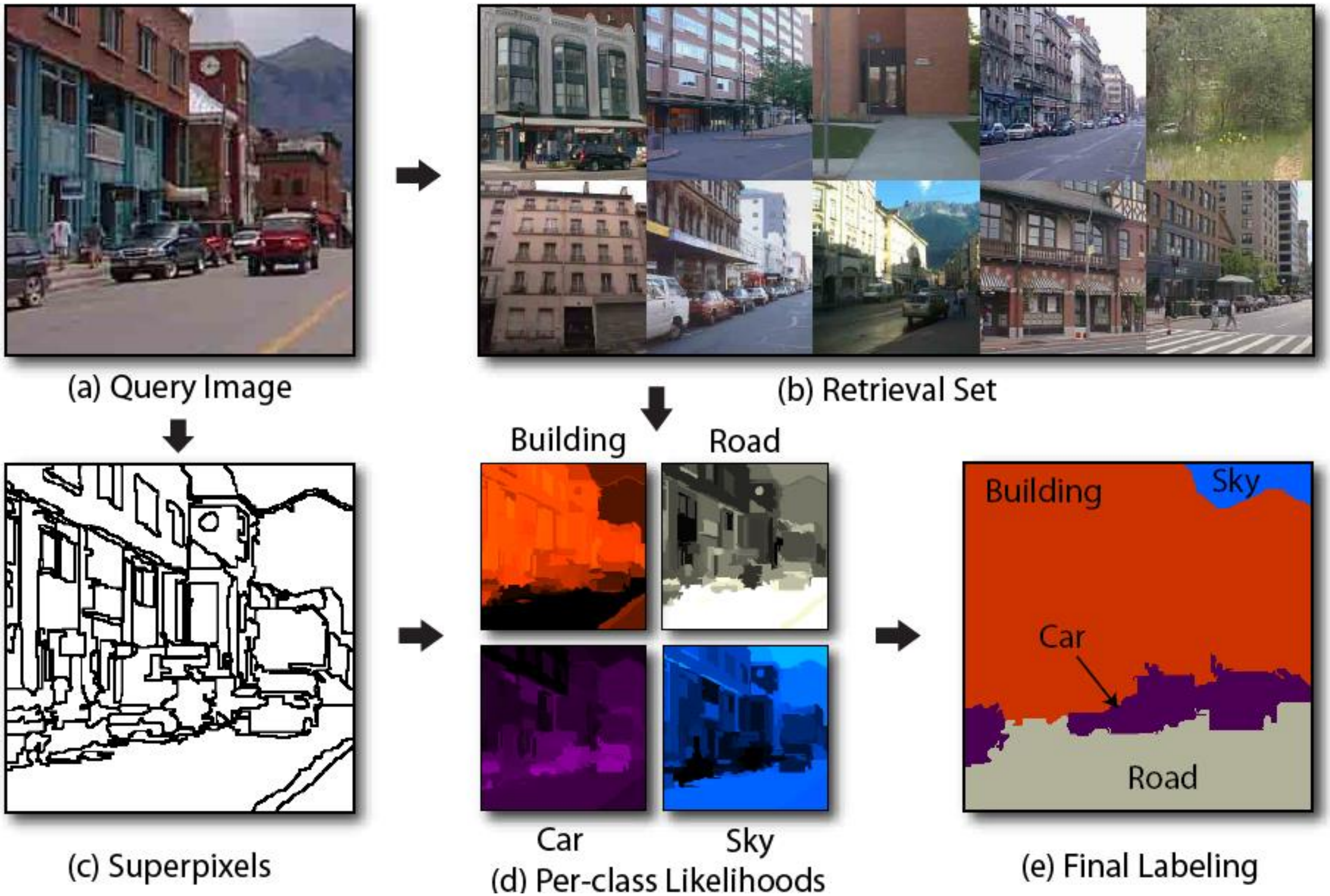
http://people.csail.mit.edu/torralba/code/spatialenvelope/

Data-driven methods



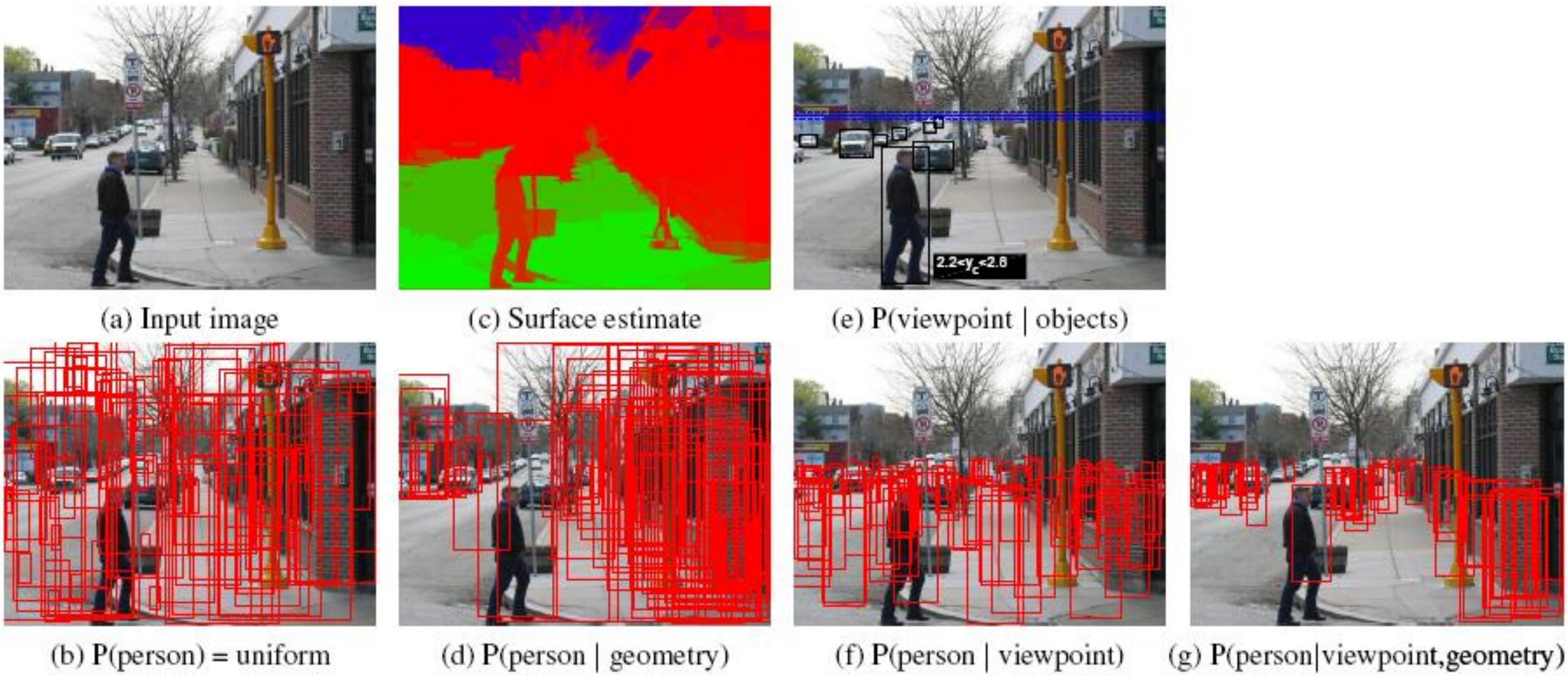
J. Hays and A. Efros, Scene Completion using Millions of Photographs, SIGGRAPH 2007

Data-driven methods



J. Tighe and S. Lazebnik, ECCV 2010

Geometric context



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

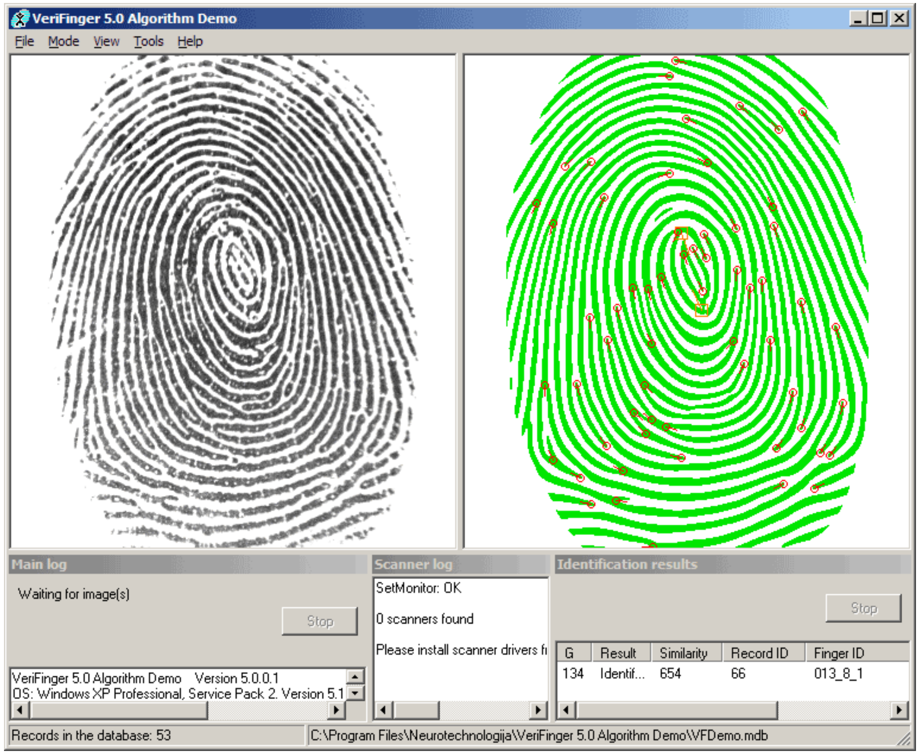
What "works" today

- Reading license plates, zip codes, checks

Svetlana Lazebnik

What "works" today

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



Svetlana Lazebnik

What “works” today

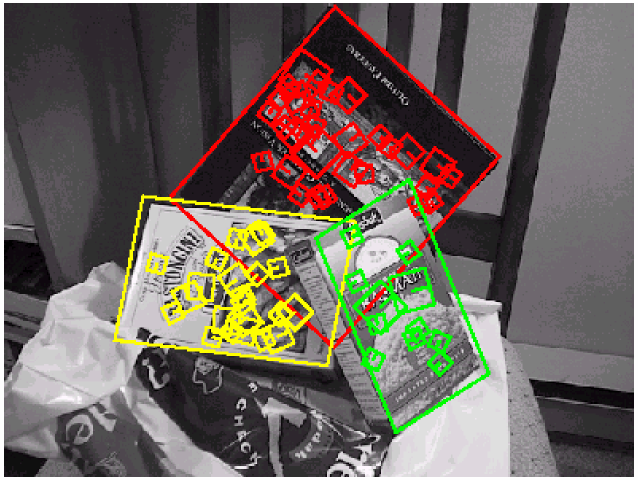
- Reading license plates, zip codes, checks
- Fingerprint recognition
- Face detection

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

Svetlana Lazebnik

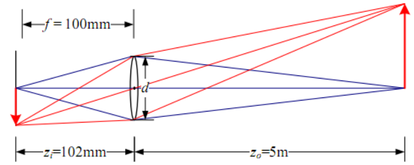
What “works” today

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



Svetlana Lazebnik

Course Outline



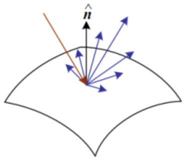


Image Formation and Processing

Light, Shape and Color

The Pin-hole Camera Model, The Digital Camera

Linear filtering, Template Matching, Image Pyramids

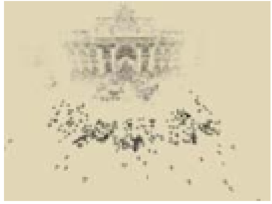
G	R	G	R
B	G	B	G
G	R	G	R
B	G	B	G

Feature Detection and Matching

Edge Detection, Interest Points: Corners and Blobs

Local Image Descriptors

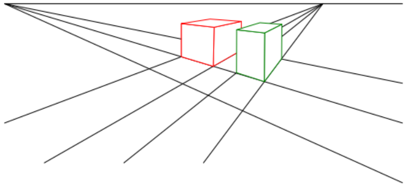
Feature Matching and Hough Transform



Multiple Views and Motion

Geometric Transformations, Camera Calibration

Feature Tracking , Stereo Vision



Segmentation and Grouping

Segmentation by Clustering, Region Merging and Growing

Advanced Methods Overview: Active Contours, Level-Sets, Graph-Theoretic Methods

Detection and Recognition

Problems and Architectures Overview

Statistical Classifiers, Bag-of-Words Model, Detection by Sliding Windows

History of Ideas in Recognition

Resources

Books

R. Szeliski, Computer Vision: Algorithms and Applications, 2010 – *available online*

D. A. Forsyth and J. Ponce, Computer Vision: A Modern Approach, 2003

L. G. Shapiro and G. C. Stockman, Computer Vision, 2001

Web

CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision

http://homepages.inf.ed.ac.uk/rbf/CVonline/

Dictionary of Computer Vision and Image Processing

http://homepages.inf.ed.ac.uk/rbf/CVDICT/

Computer Vision Online

http://www.computervisiononline.com/

Programming

Development environments/languages: Matlab, Python and C/C++

Toolboxes and APIs: OpenCV, VLFeat Matlab Toolbox, Piotr's Computer Vision Matlab Toolbox, EasyCamCalib Software, FLANN, Point Cloud Library PCL, LibSVM, Camera Calibration Toolbox for Matlab