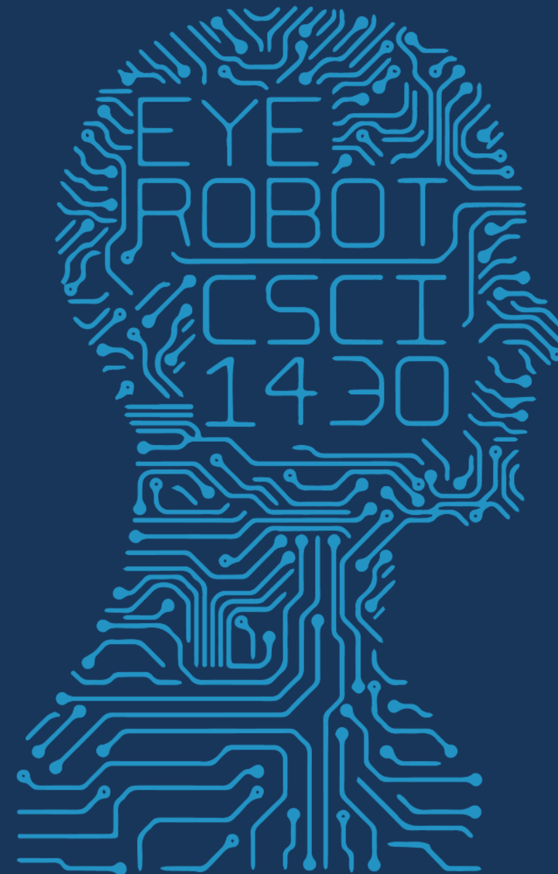
1950

Future Vision

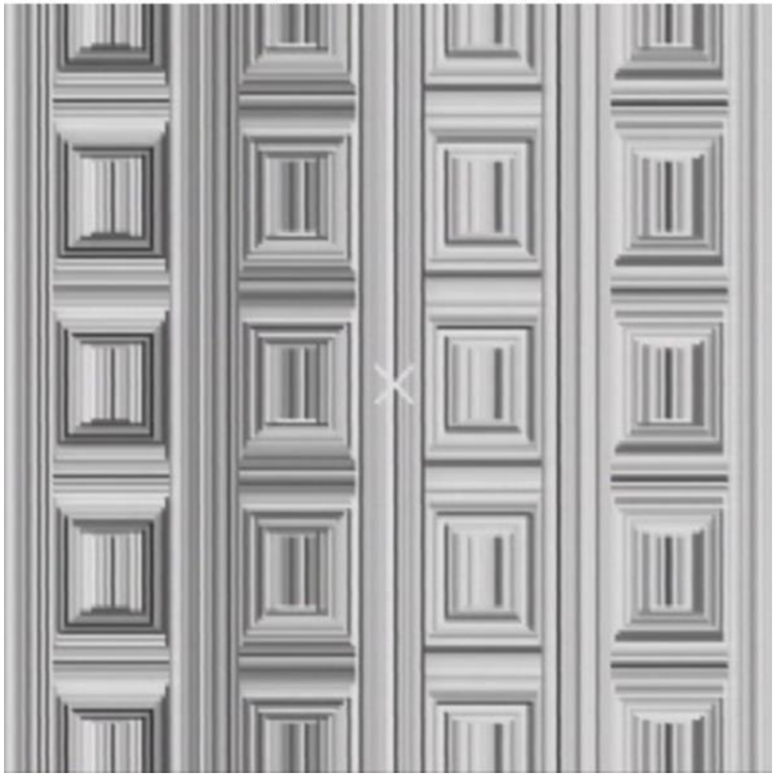


2017 MWF 1pm

Computer Vision

Coffer Illusion

How many circles do you see?



Coffer Illusion

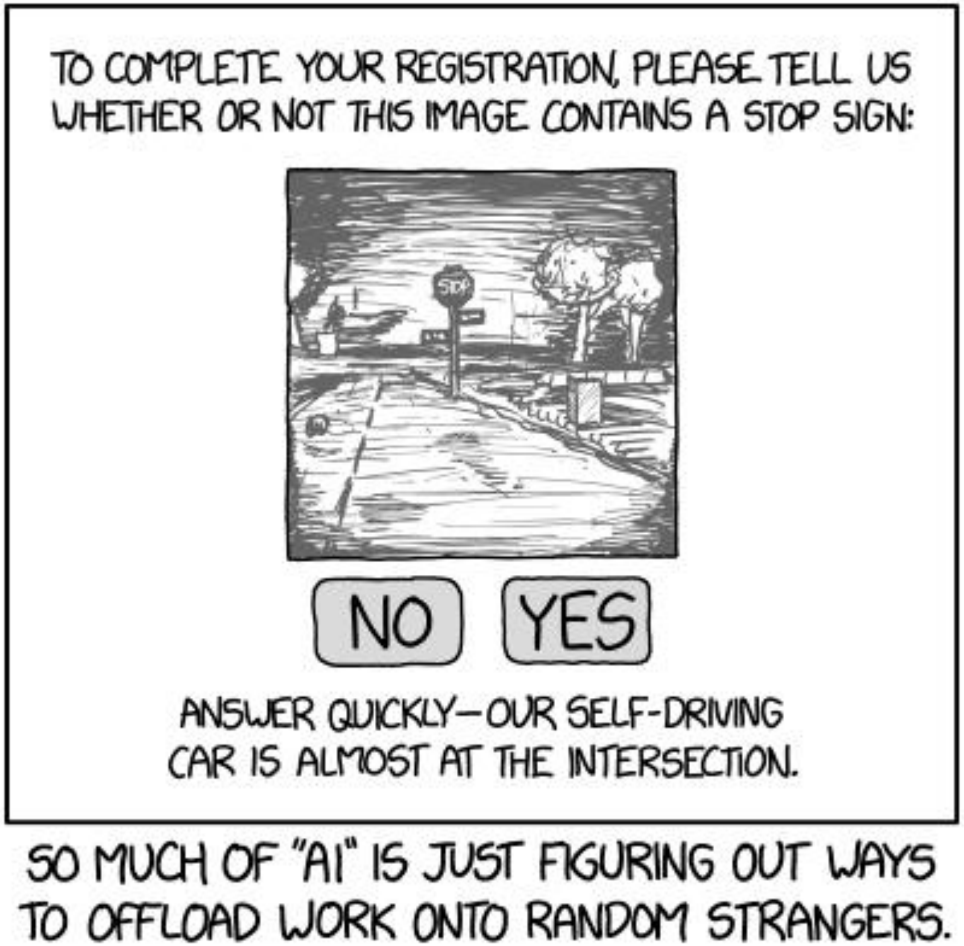
An elephant standing on top of a basket being held by a woman

MS COCO

wordseye.com

Thank you Trent Green

Thanks to **Iuliu Balibanu**



Alt-text: "Crowdsourced steering" doesn't sound quite as appealing as "self driving".

Machine Learning Problems

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Supervised learning

$$f(\mathbf{x}) = y$$

Prediction function — Image feature — Output (label)

Training: Given a *training set* of labeled examples:

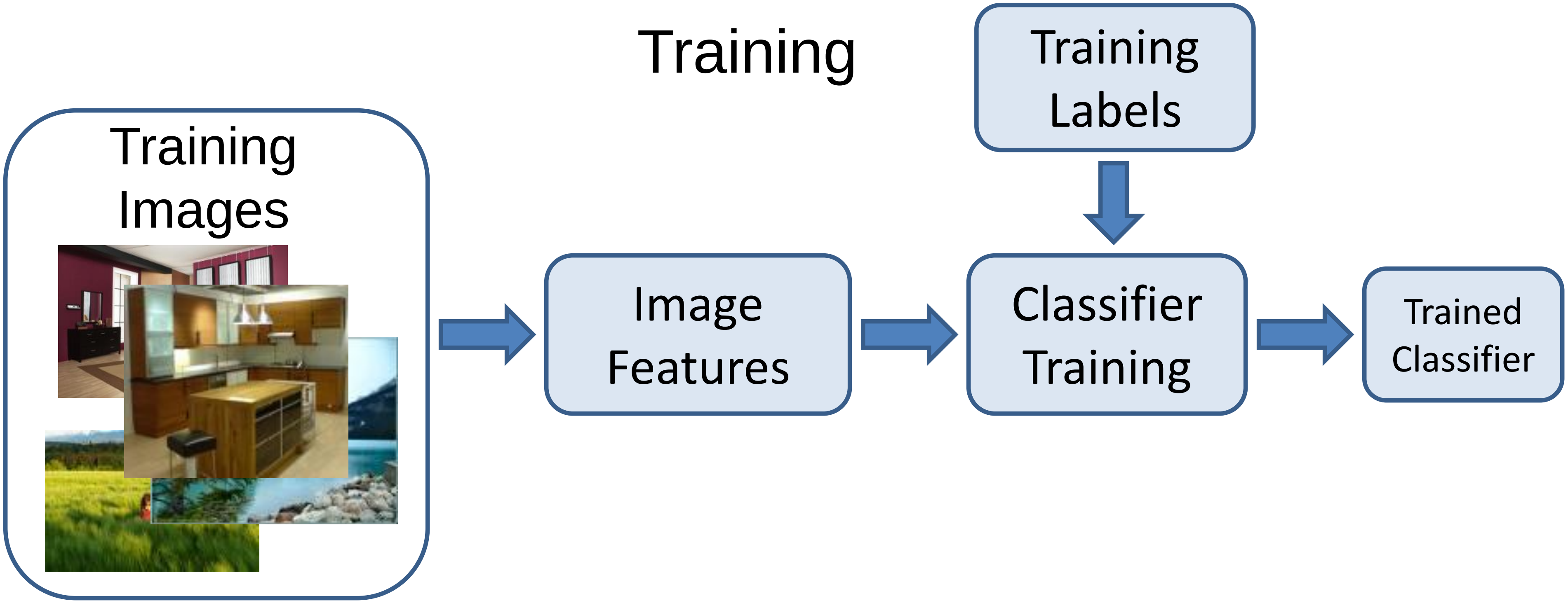
$$\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$$

Estimate the prediction function f by minimizing the prediction error on the training set.

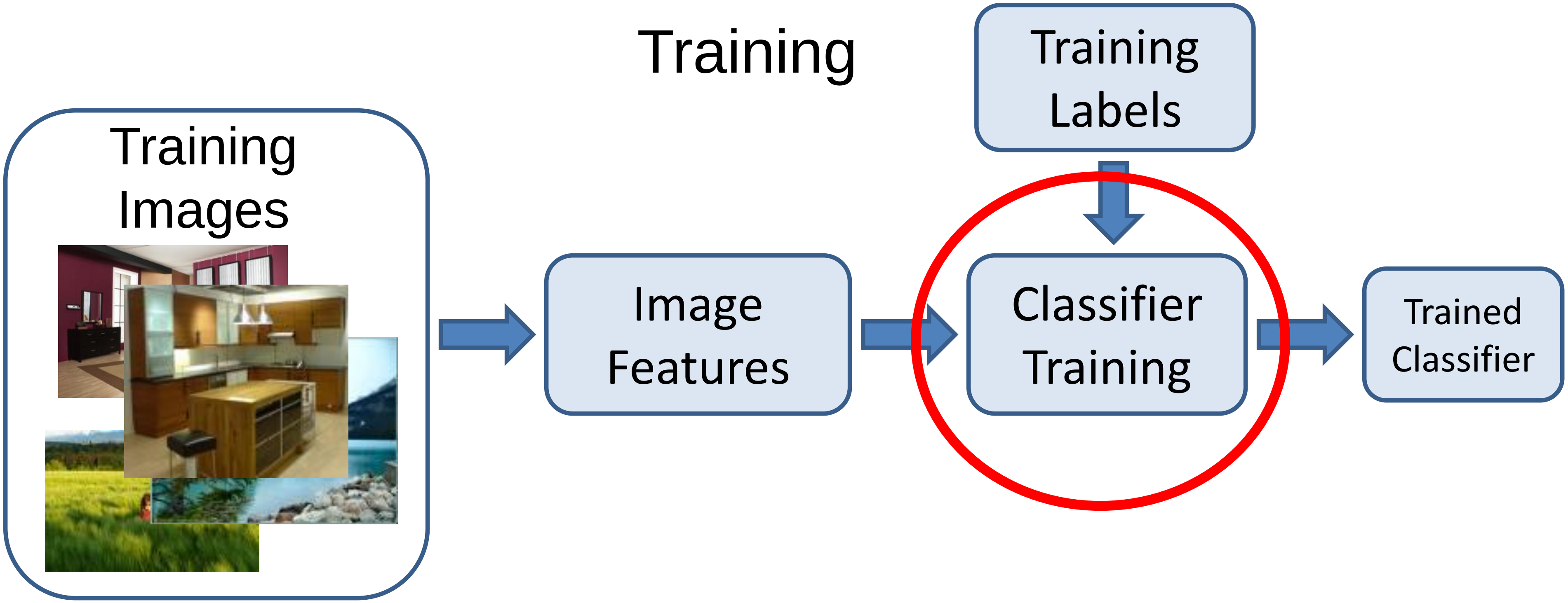
Testing: Apply f to a unseen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$ to *classify* $\mathbf{x}$.

Slide credit: L. Lazebnik

Image Categorization



Classifiers



Learning a classifier

Given a set of features with corresponding labels, learn a function to predict the labels from the features.

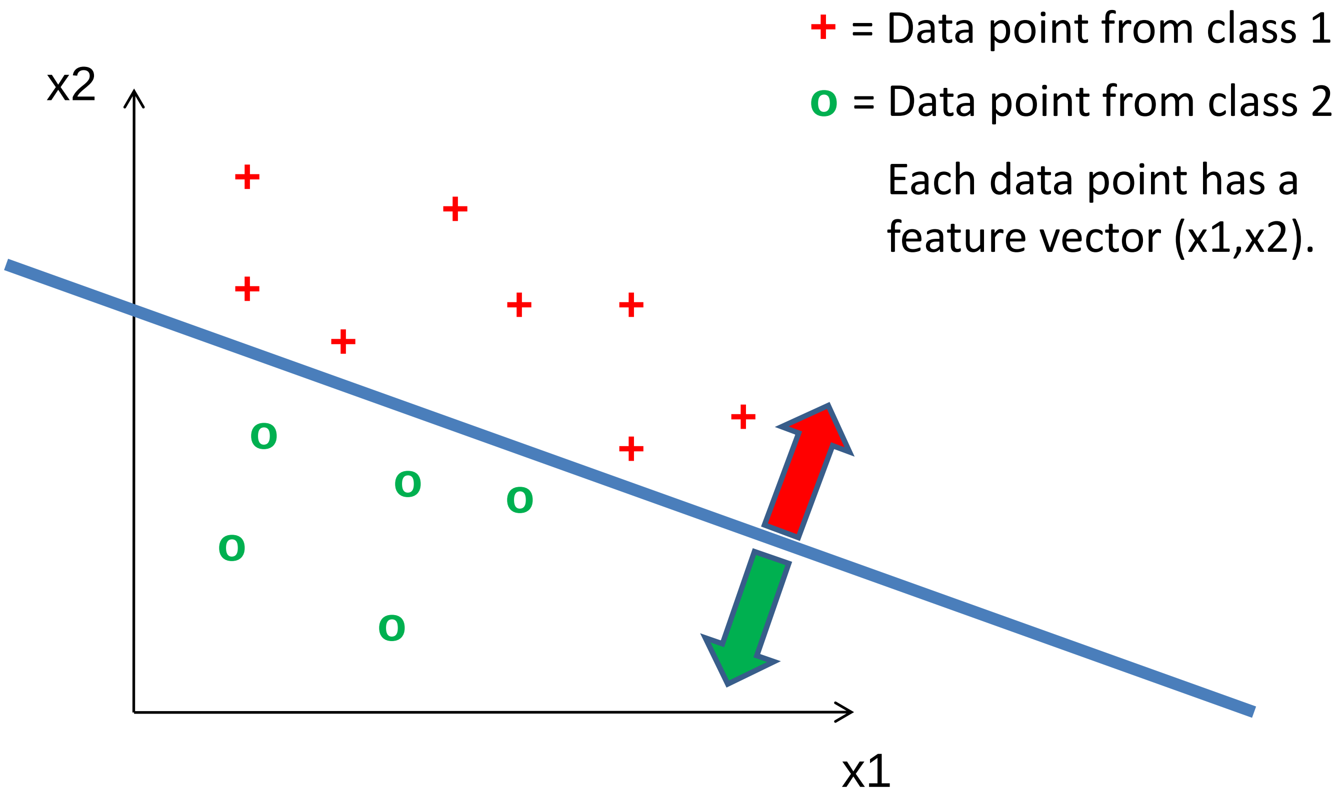
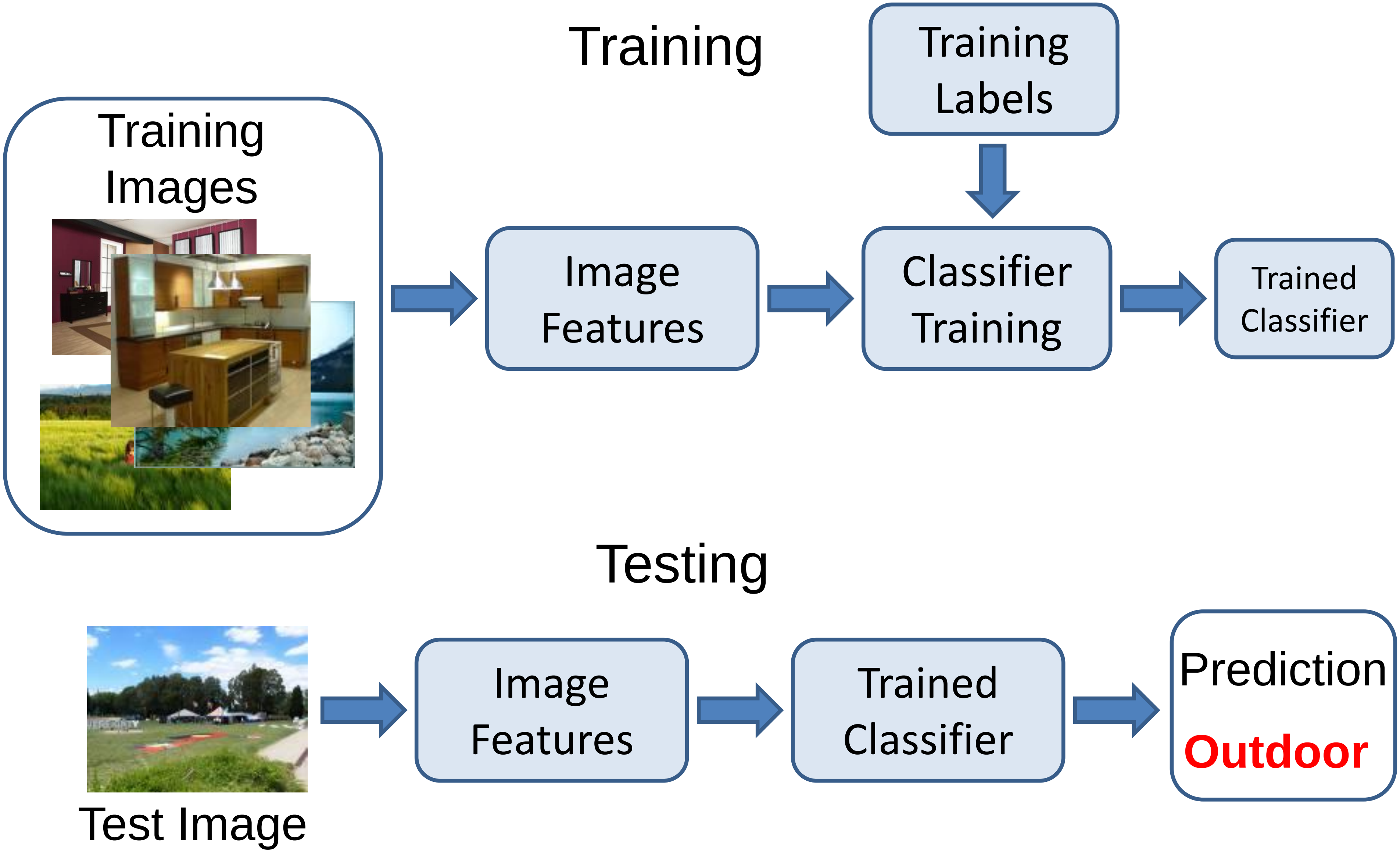


Image Categorization



Example: Scene Categorization

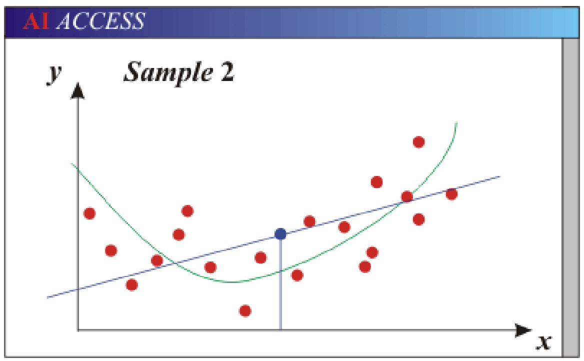
- Is this a kitchen?

Bias-Variance Trade-off

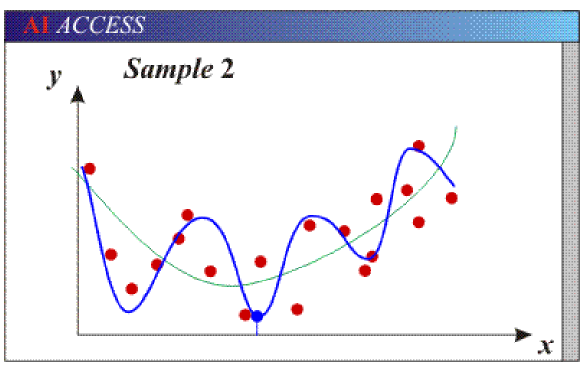
Bias: *error in model assumptions*; how much the average model over all training sets differs from the true model.

Variance: how much models estimated from different training sets differ from each other.



Models with too few parameters are inaccurate because of a large bias.

- Not enough flexibility!



Models with too many parameters are inaccurate because of a large variance.

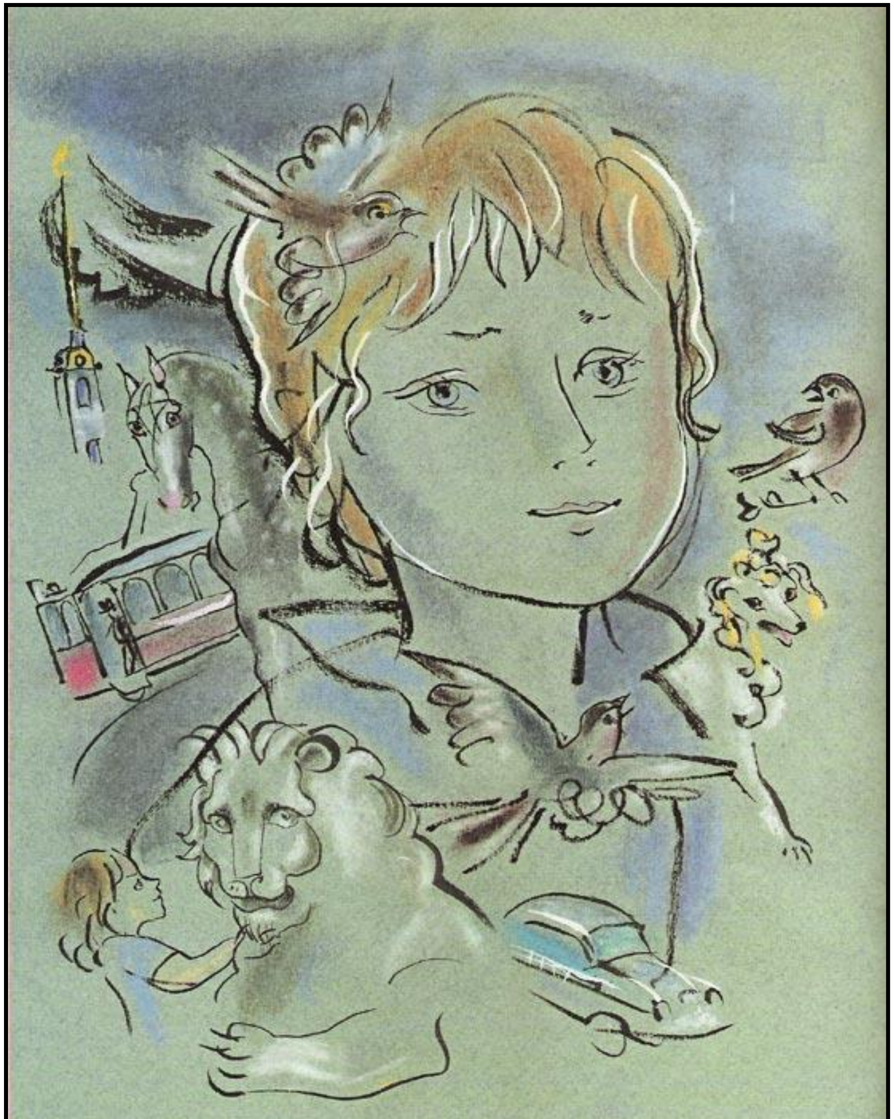
- Too much sensitivity to the sample.

ML crash course

Nice write-up of the bias-variance issues

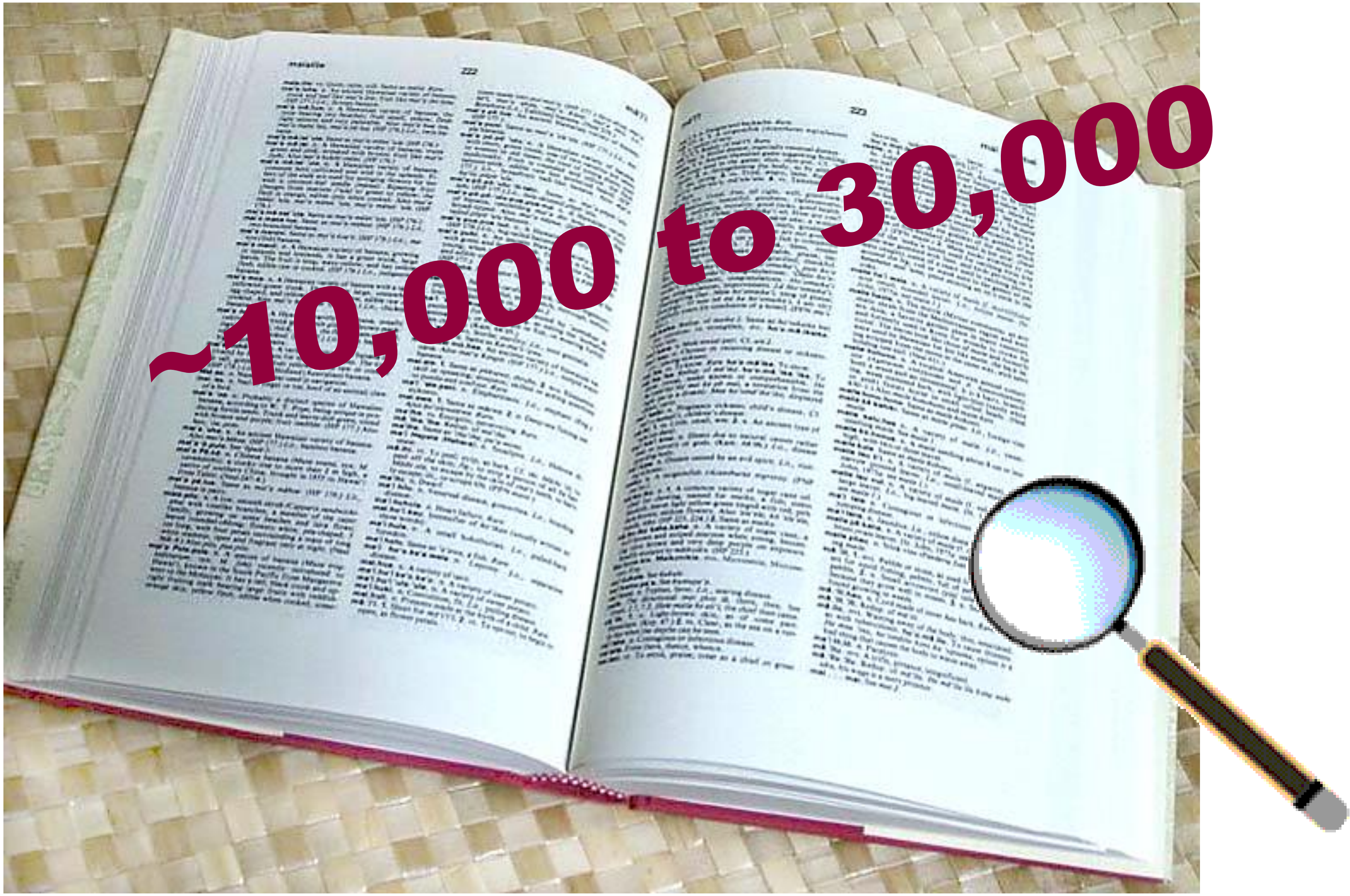
http://www.learnopencv.com/bias-variance-tradeoff-in-machine-learning/

Recognition: Overview and History



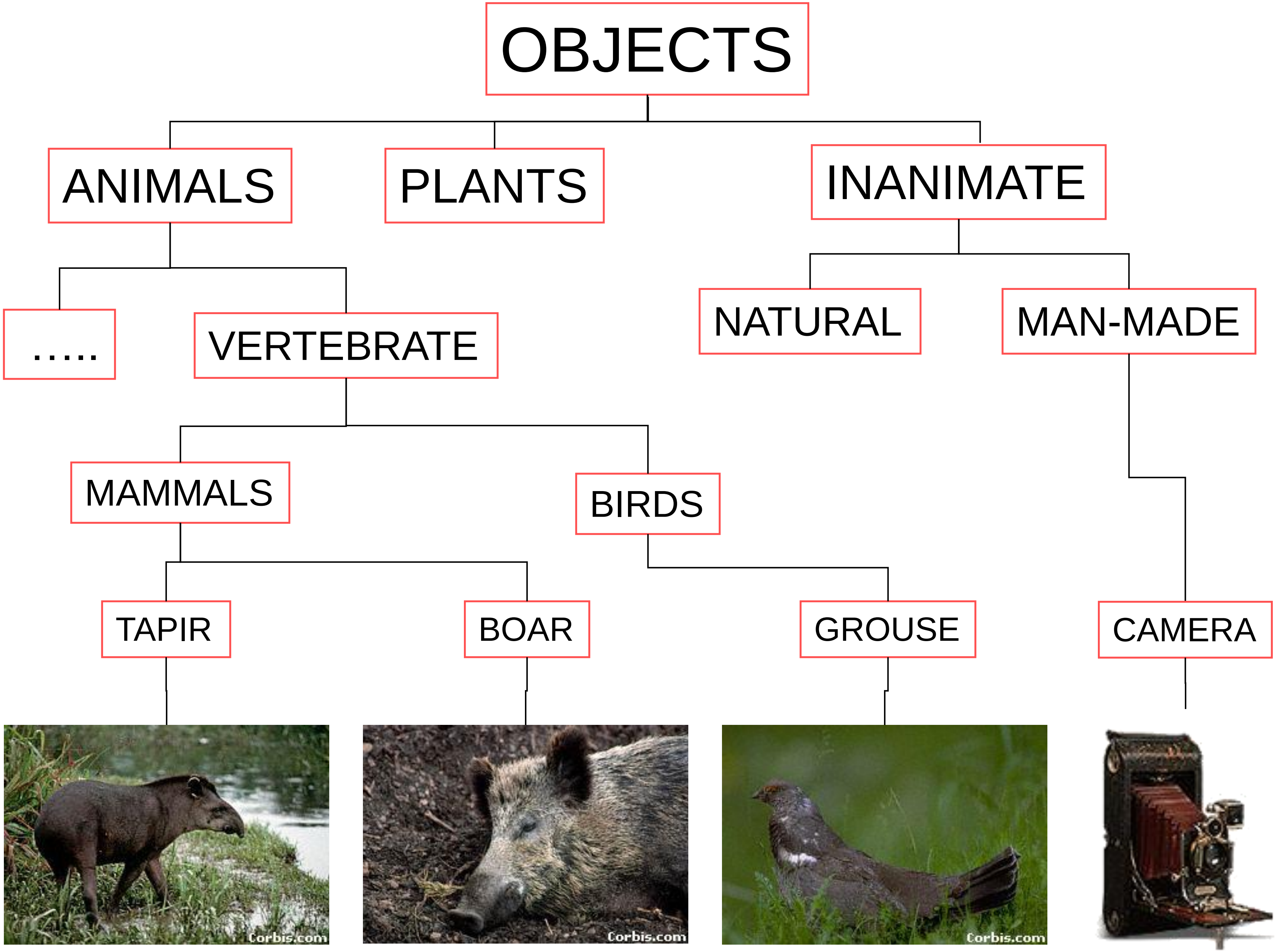
Slides from James Hays, Lana Lazebnik, Fei-Fei Li, Rob Fergus, Antonio Torralba, and Jean Ponce

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

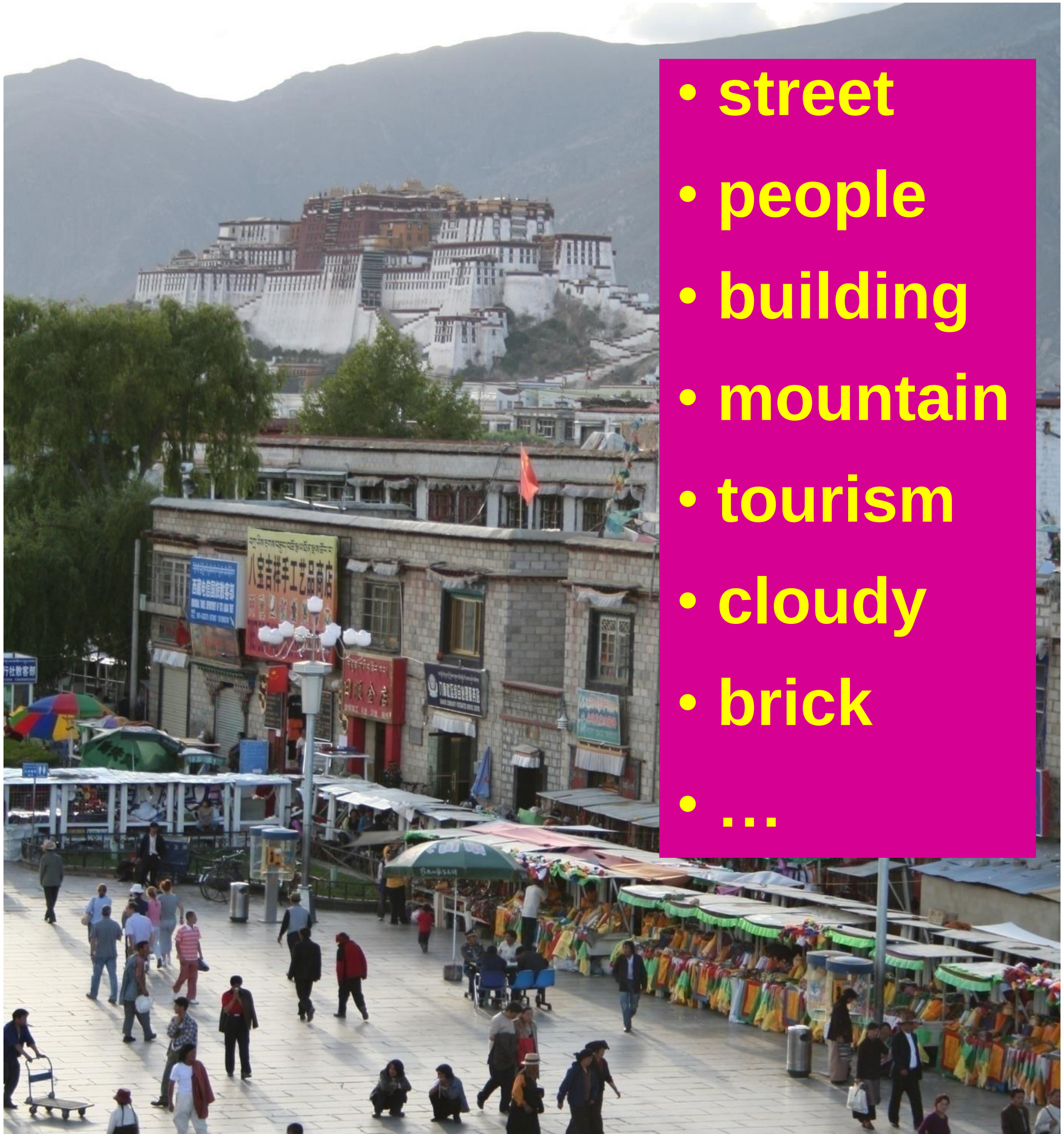
Svetlana Lazebnik

Scene categorization or classification

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

Svetlana Lazebnik

Image annotation / tagging / attributes

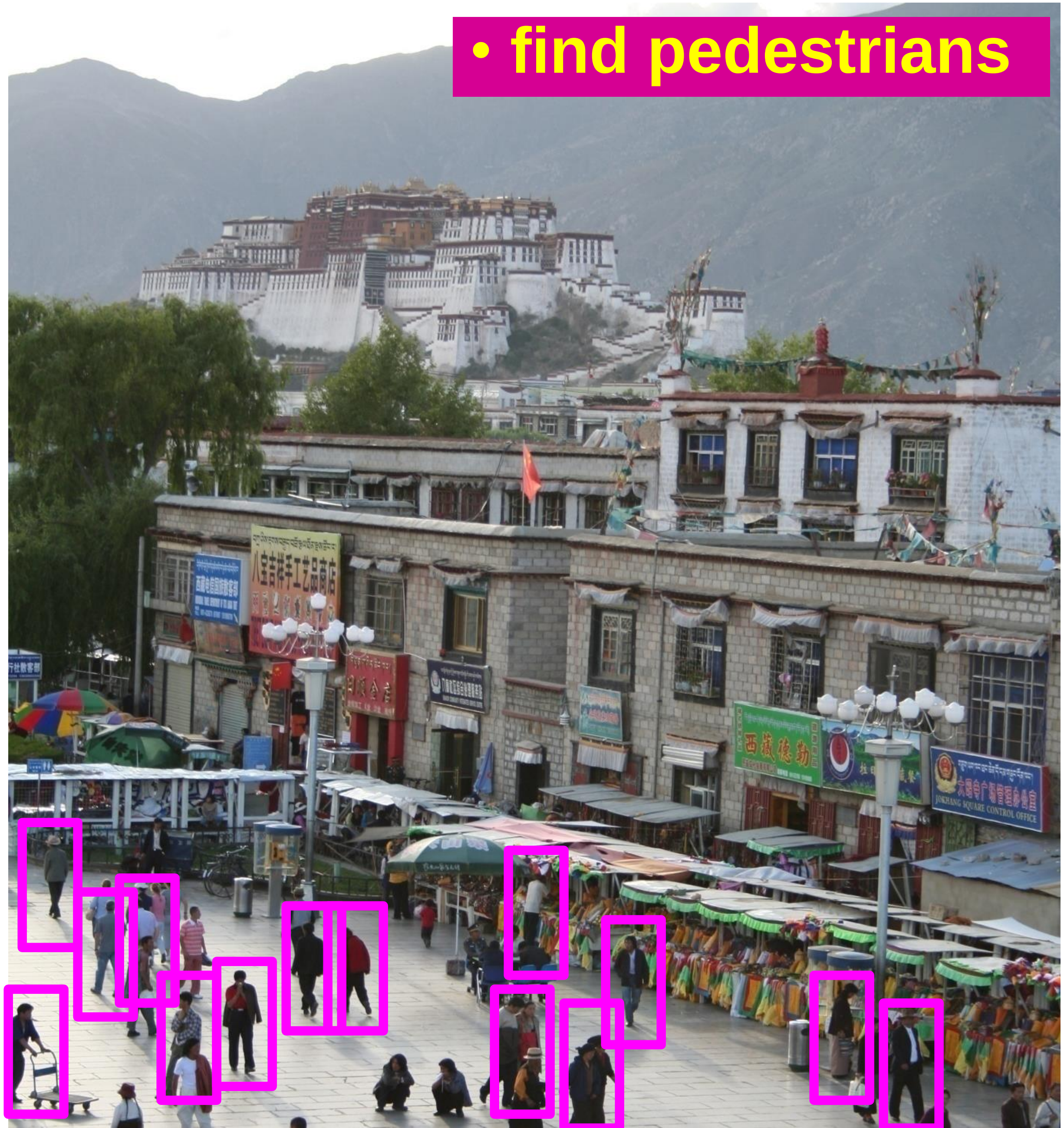


- **street**
- **people**
- **building**
- **mountain**
- **tourism**
- **cloudy**
- **brick**
- **…**

Svetlana Lazebnik

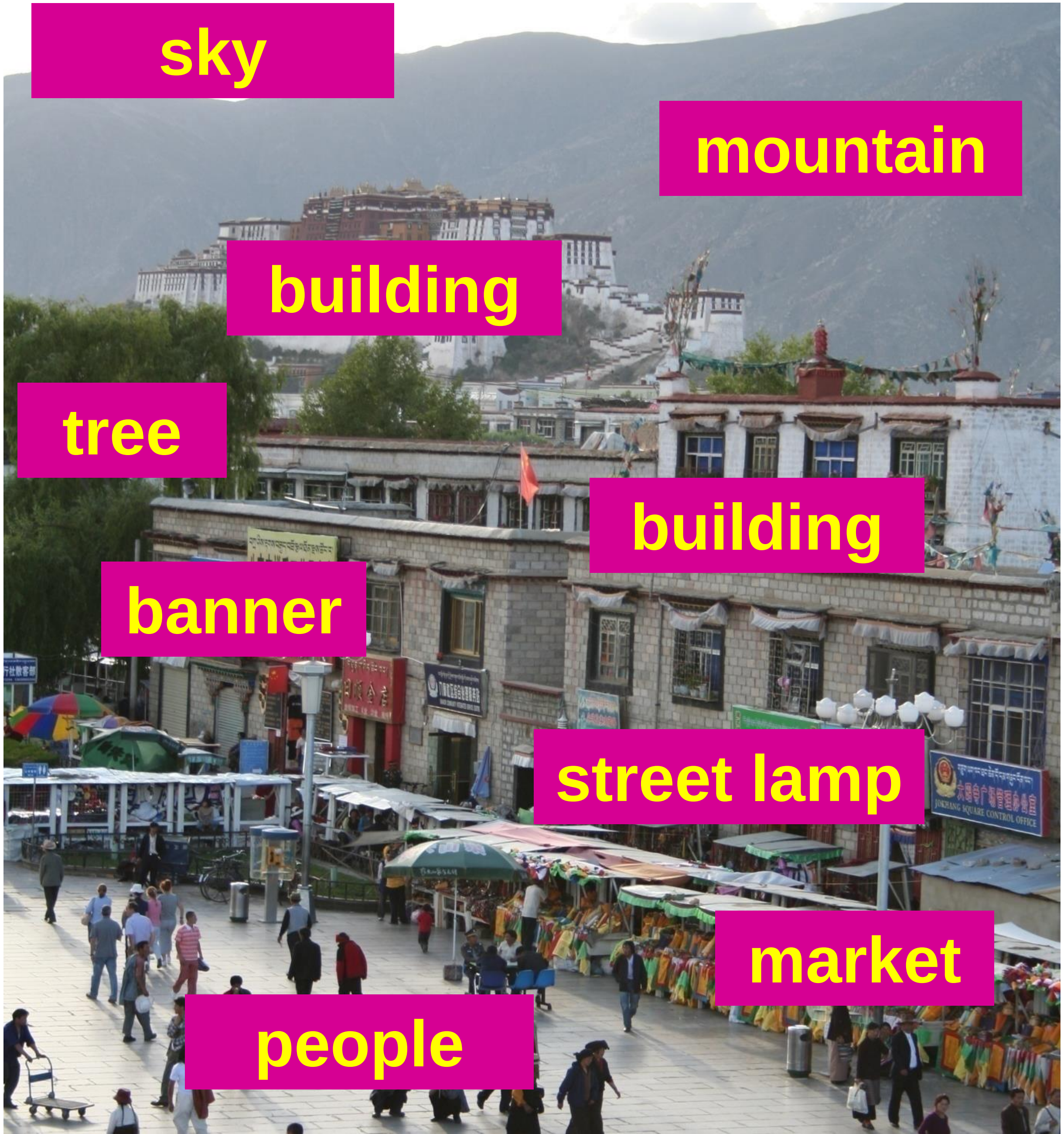
Object detection

- **find pedestrians**



Svetlana Lazebnik

Image parsing / semantic segmentation

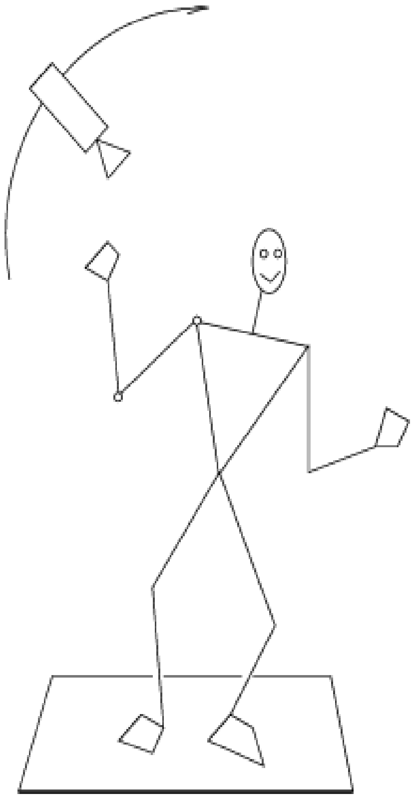


Svetlana Lazebnik

Scene understanding?

Svetlana Lazebnik

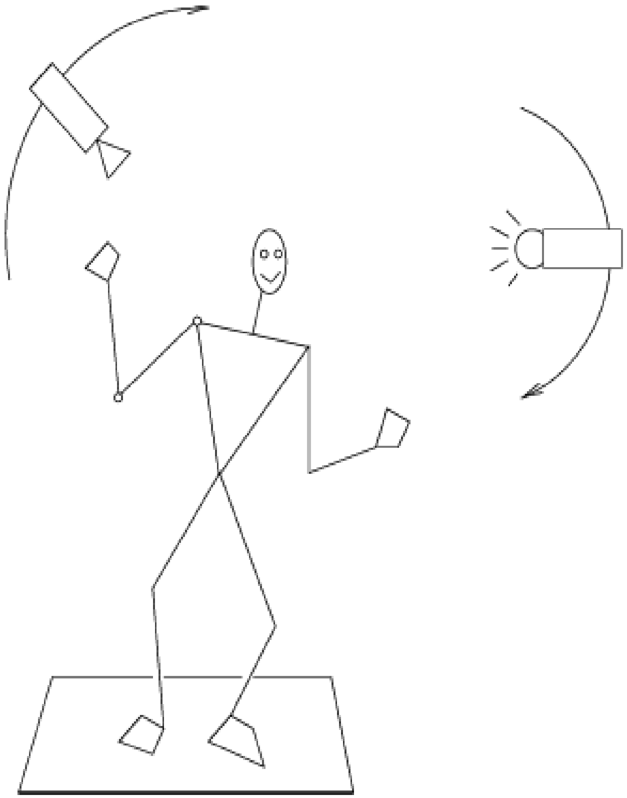
Recognition is all about modeling variability



Variability: Camera position

Svetlana Lazebnik

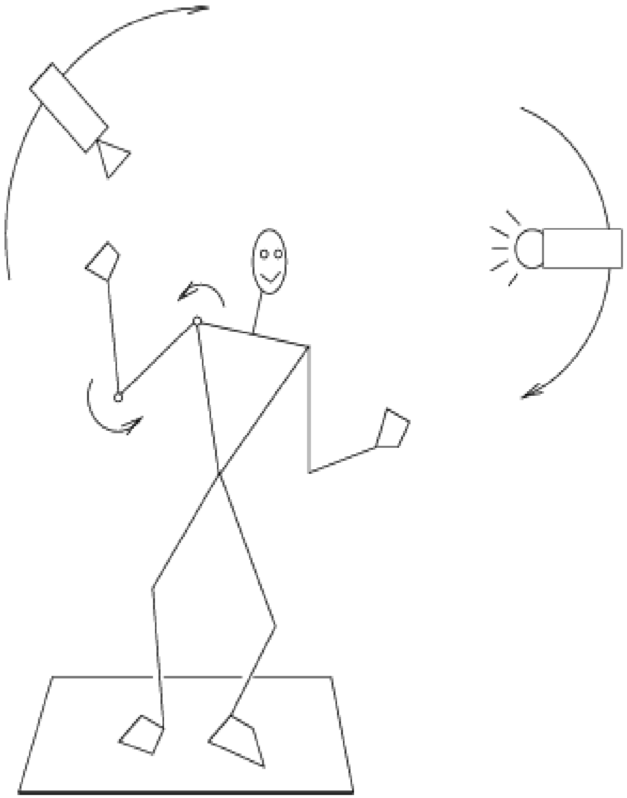
Recognition is all about modeling variability



Variability: Camera position
Illumination

Svetlana Lazebnik

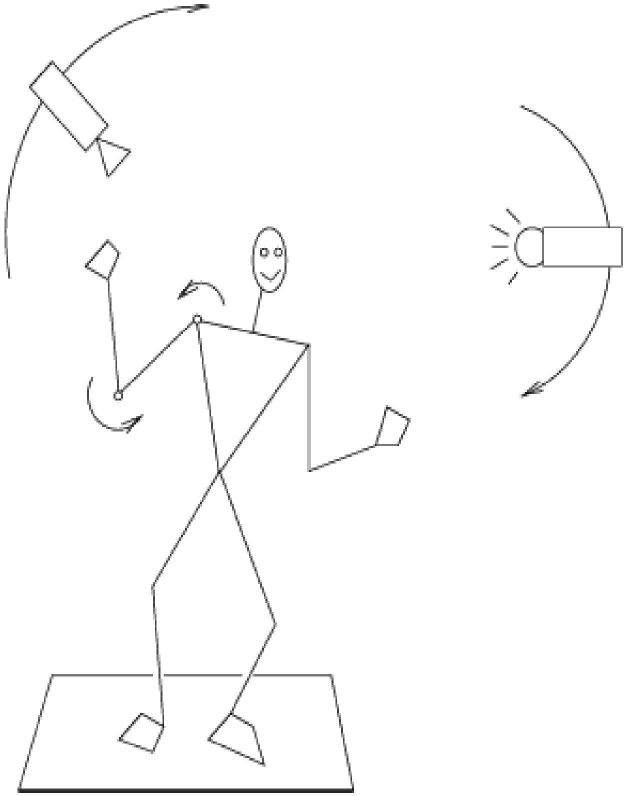
Recognition is all about modeling variability



Variability: Camera position
Illumination
Shape parameters

Svetlana Lazebnik

Recognition is all about modeling variability

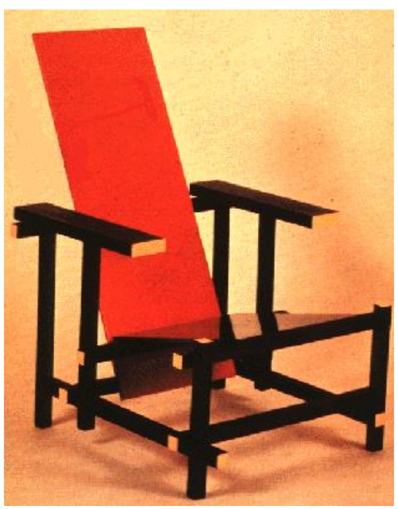


Variability: Camera position
Illumination
Shape parameters

Within-class variations?

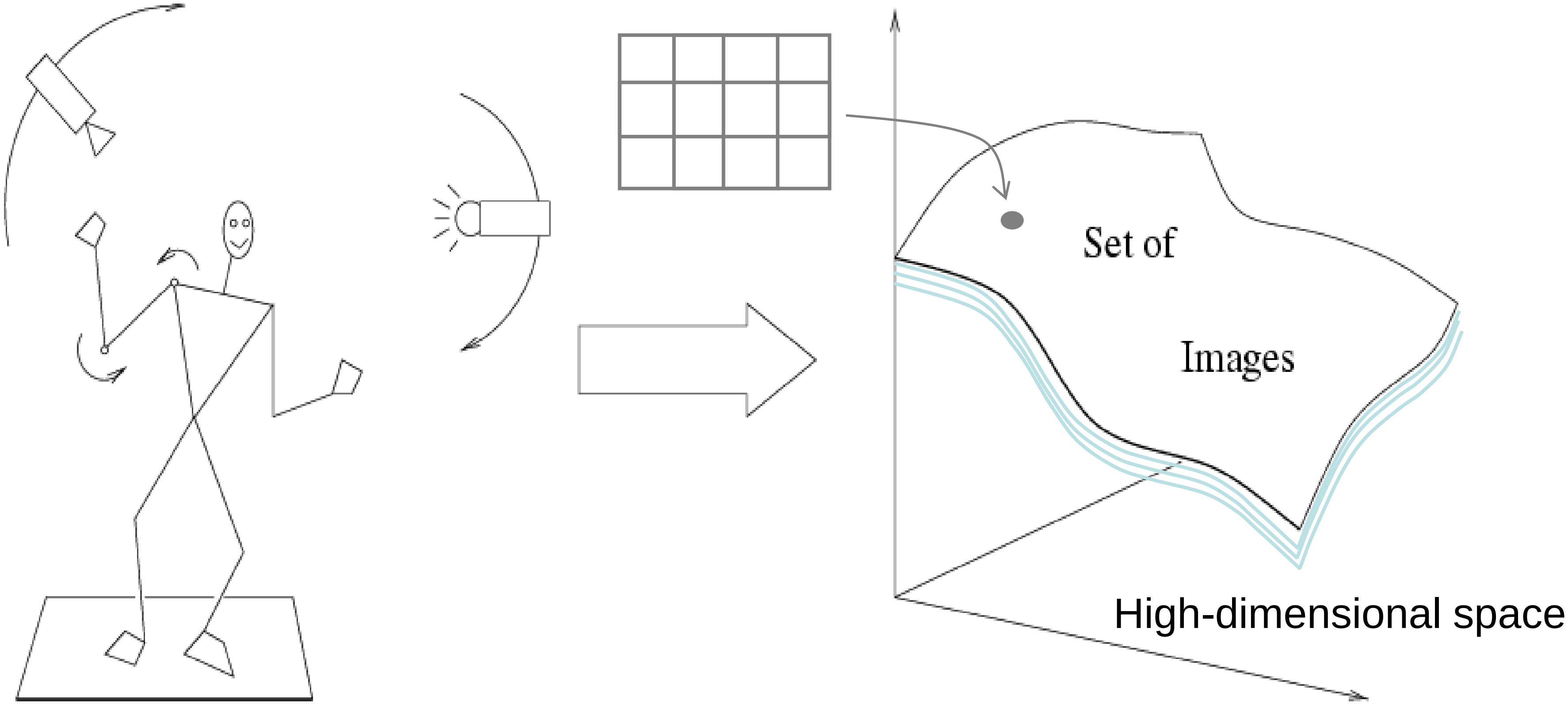
Svetlana Lazebnik

Within-class variations



Svetlana Lazebnik

Recognition is all about modeling variability



Variability: Camera position
Illumination
Shape parameters
Within-class variation

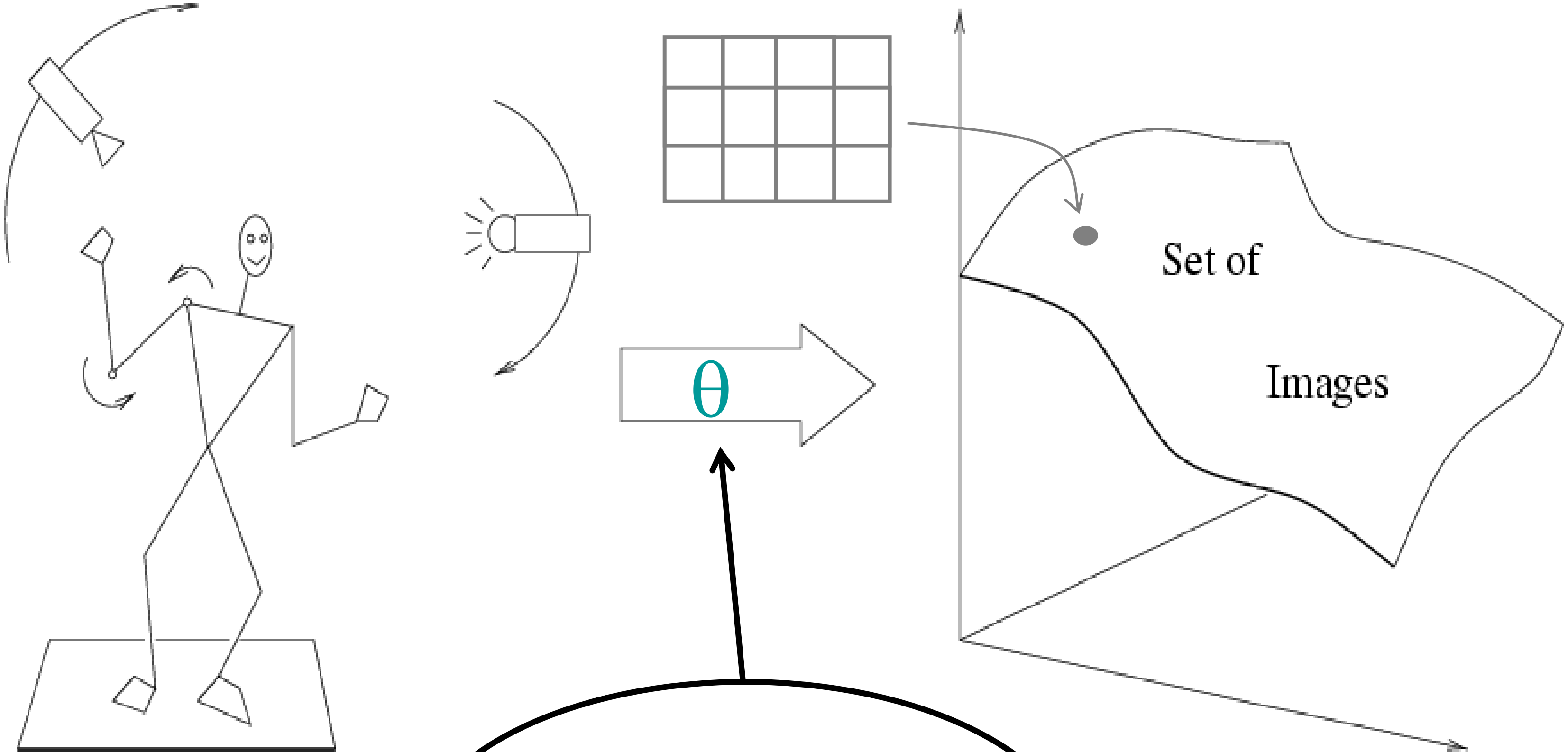
Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era

No digital cameras!
Slow compute!

Svetlana Lazebnik



Variability:

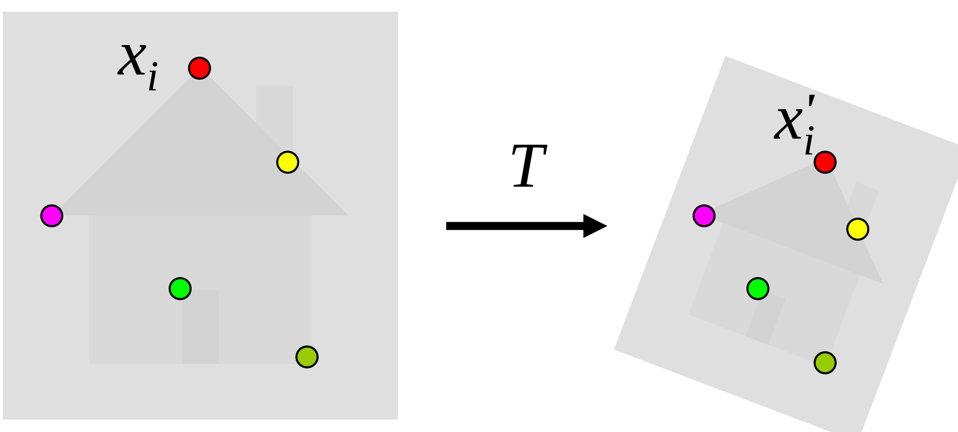
Camera position
Illumination

Shape is known

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

Alignment

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

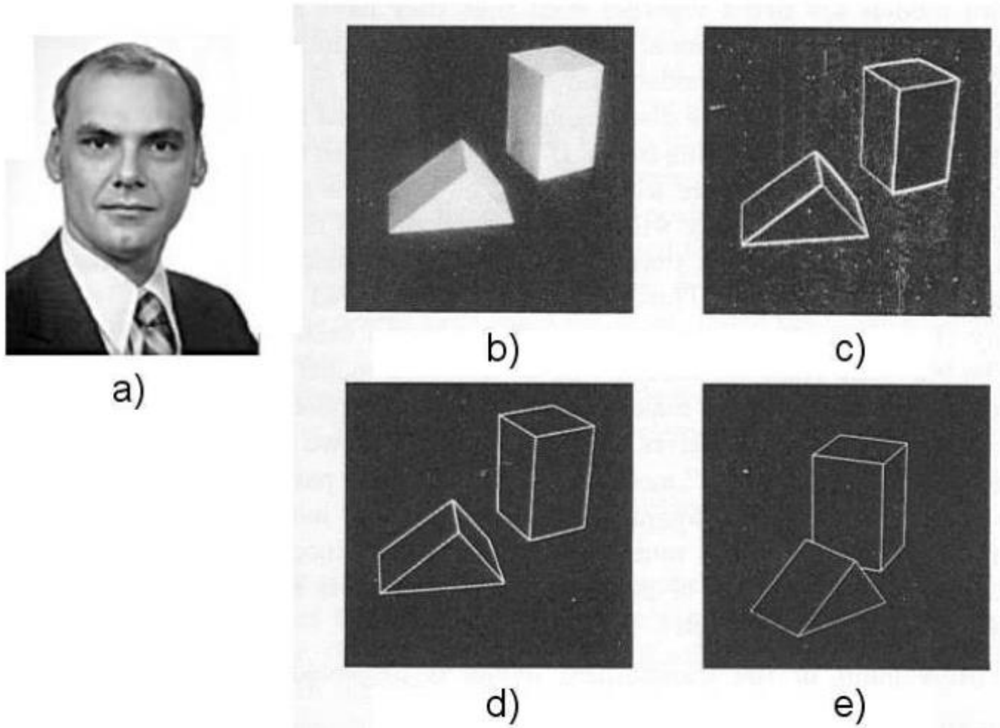


Find transformation T that minimizes

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

Svetlana Lazebnik

Recognition as an alignment problem: Block world

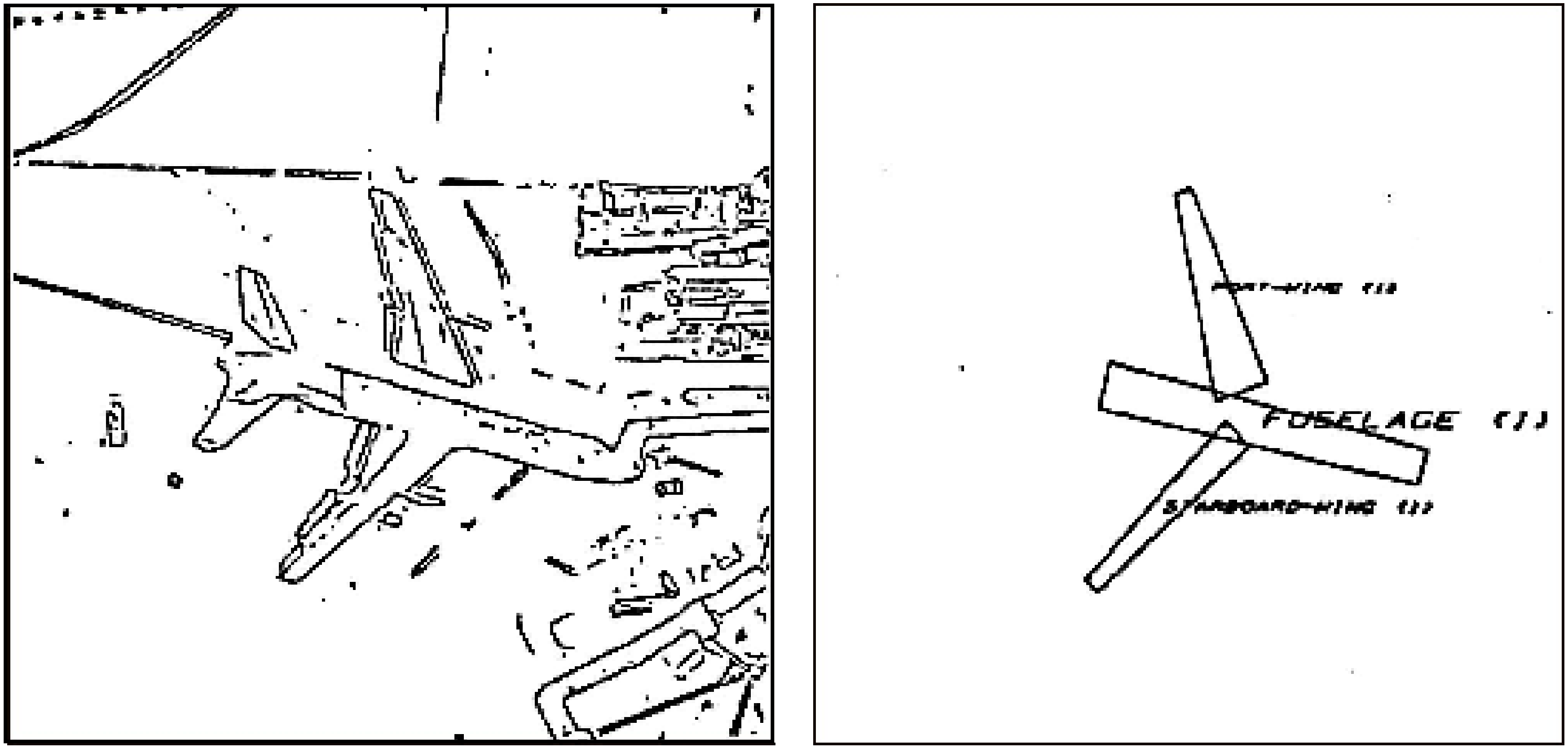


L. G. Roberts
Machine Perception of Three Dimensional Solids,
Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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

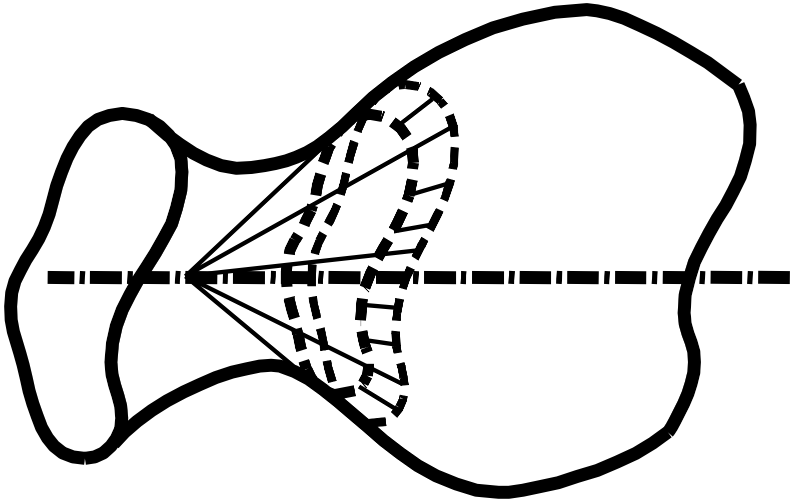
Representing and recognizing object categories is harder...



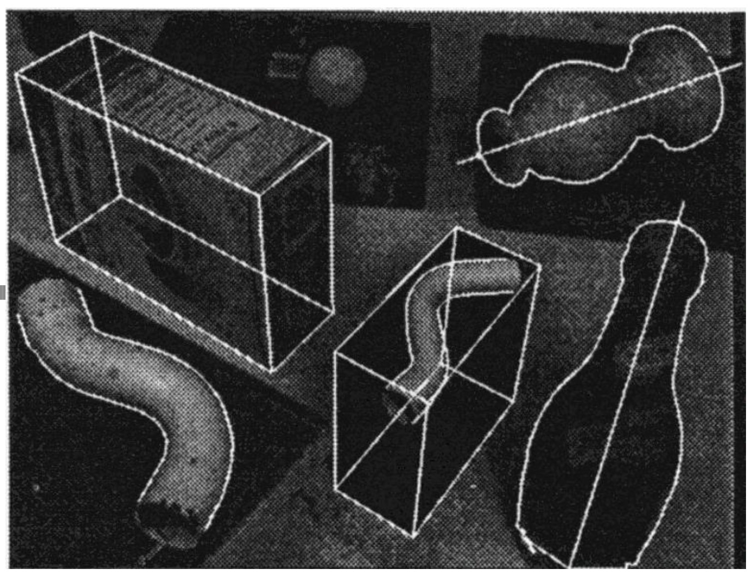
ACRONYM (Brooks and Binford, 1981)

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

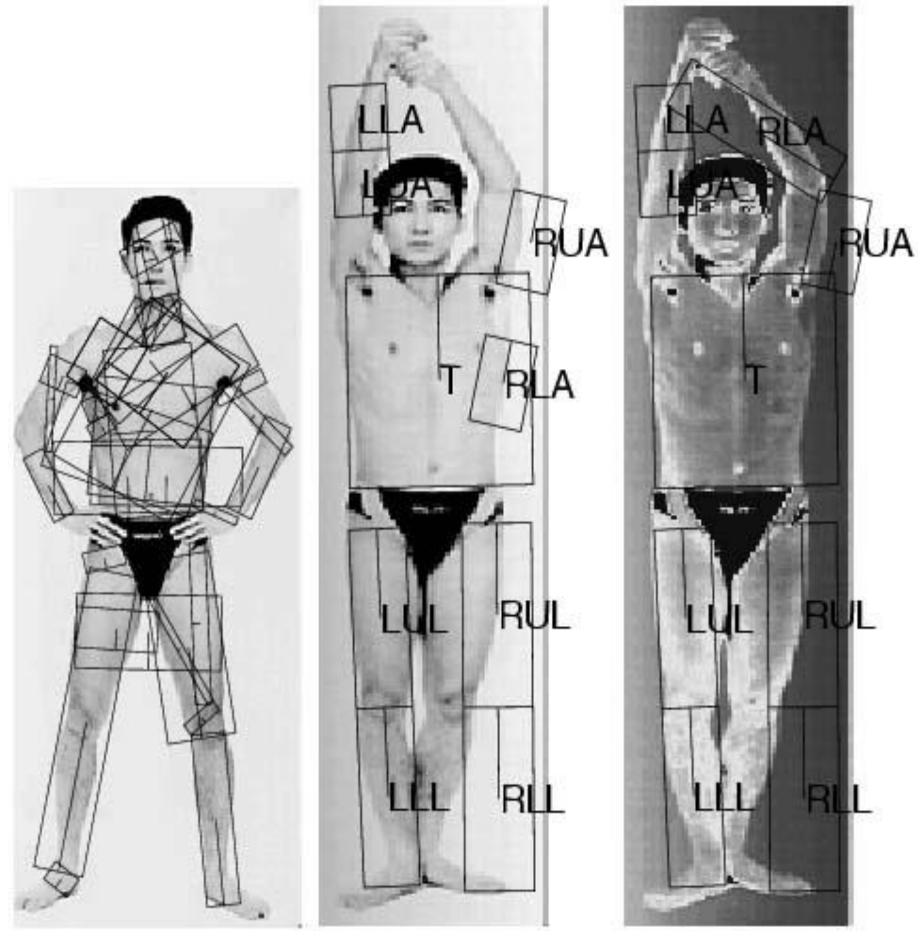
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



Zisserman et al. (1995)



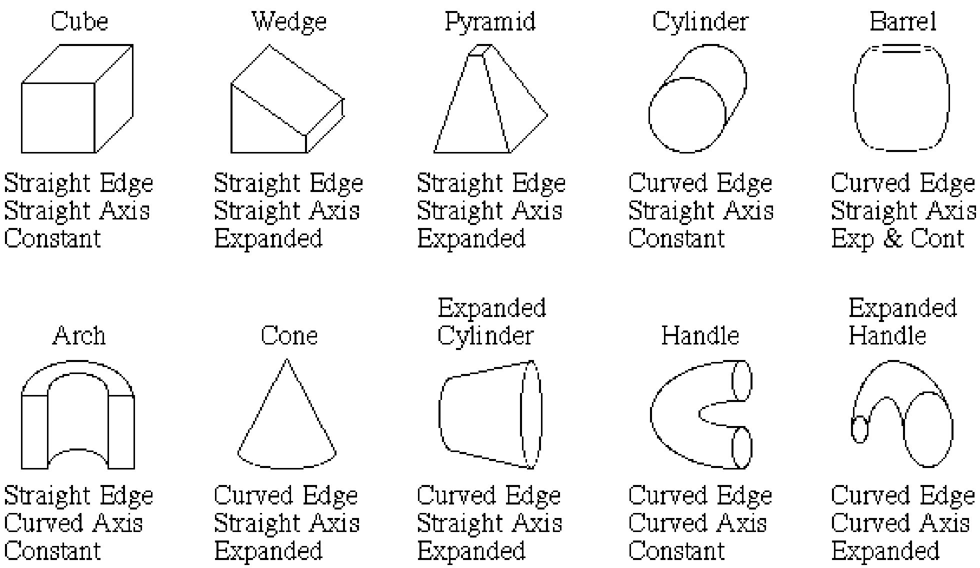
Forsyth (2000)

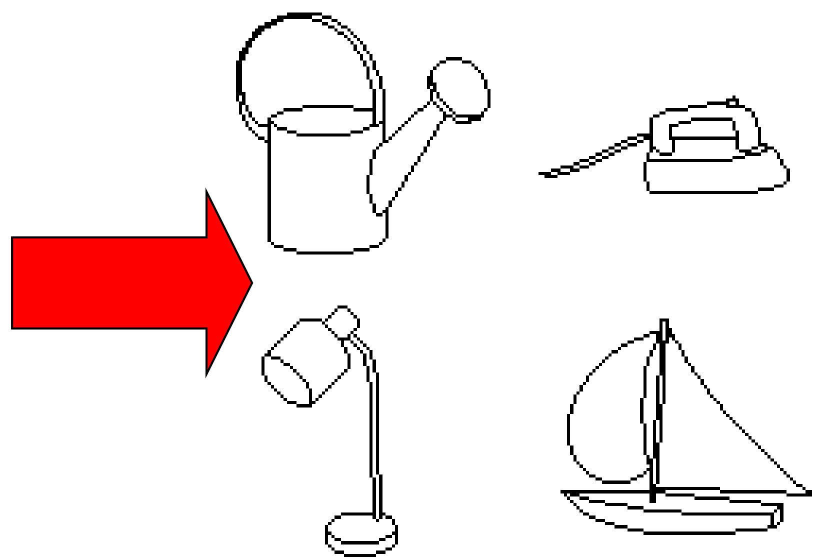
Svetlana Lazebnik

Recognition by components

Biederman (1987)

Primitives (geons) Objects





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

Svetlana Lazebnik

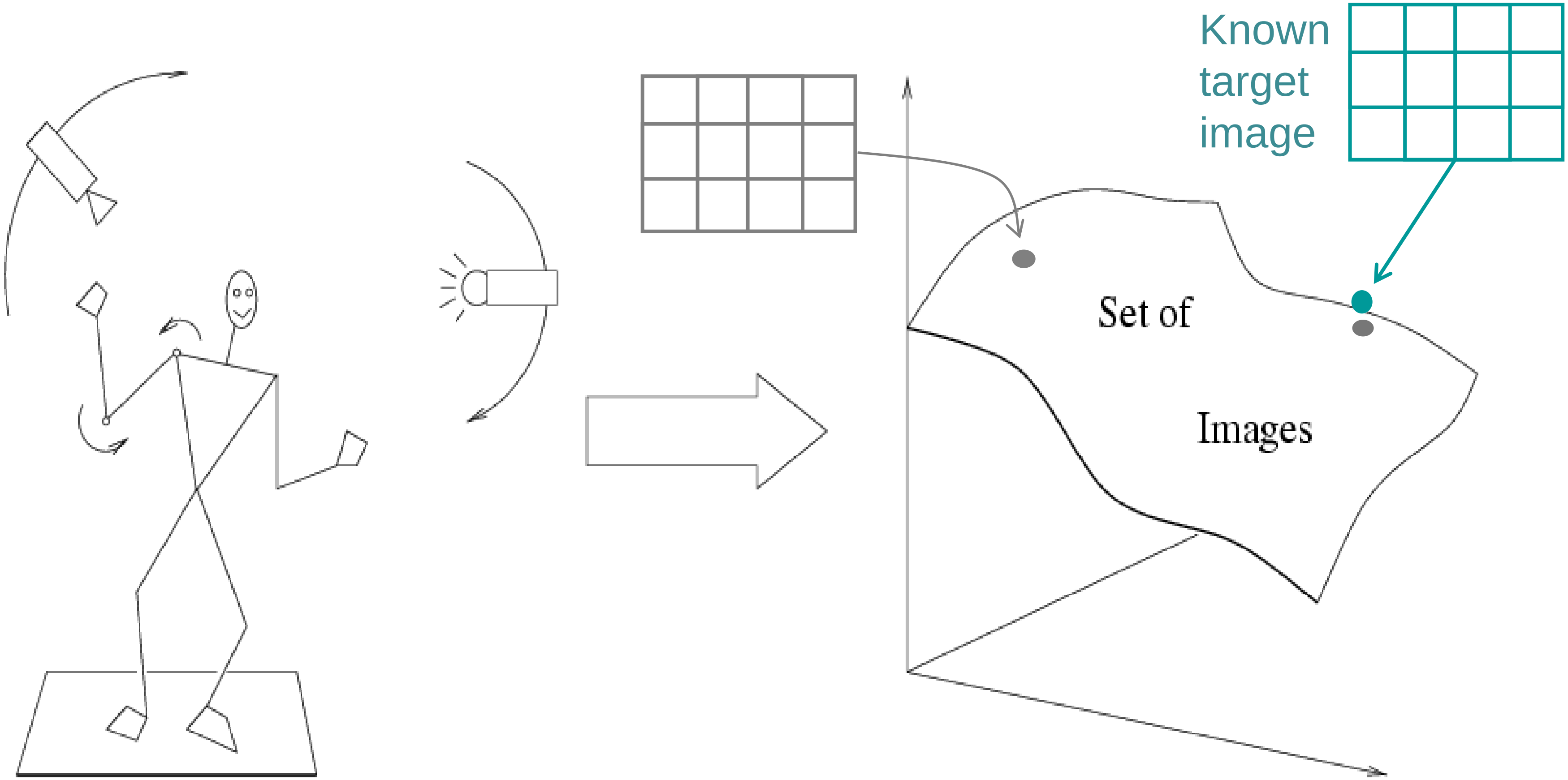
History of ideas in recognition

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

No digital cameras!
Slow compute!

Slow compute!

Svetlana Lazebnik

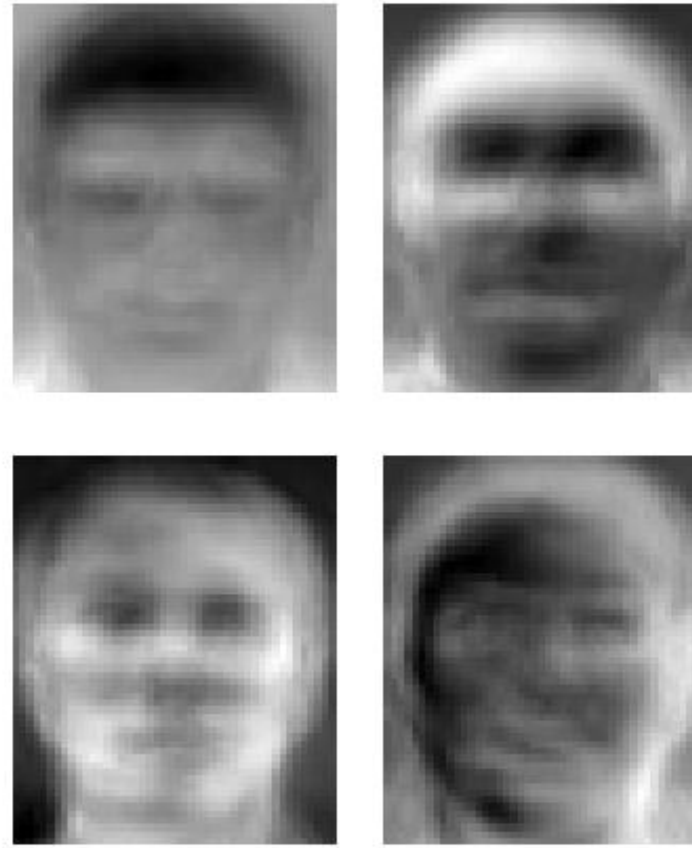


Empirical models of image variability

Appearance-based techniques

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

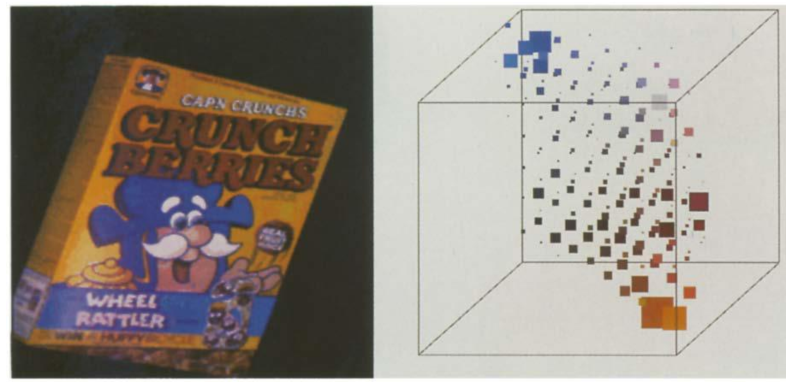
Eigenfaces (Turk & Pentland, 1991)

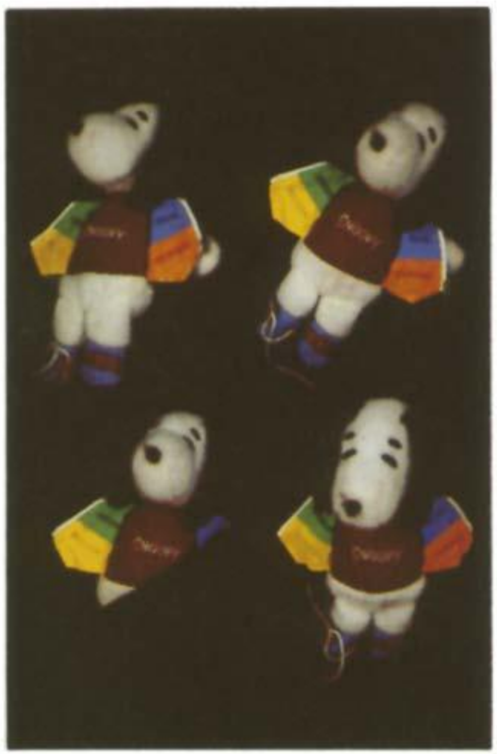


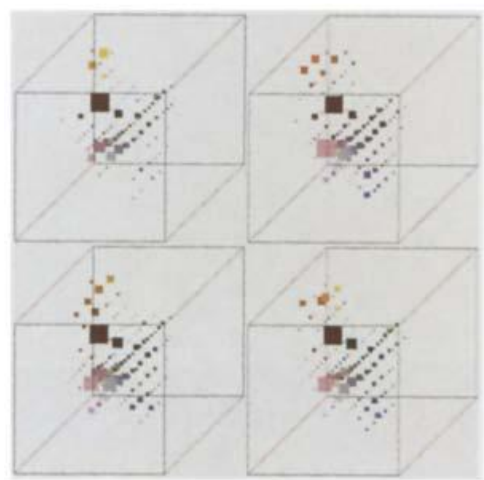
Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

Svetlana Lazebnik

Color Histograms







Swain and Ballard, Color Indexing, IJCV 1991.

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era

No digital cameras!
Slow compute!

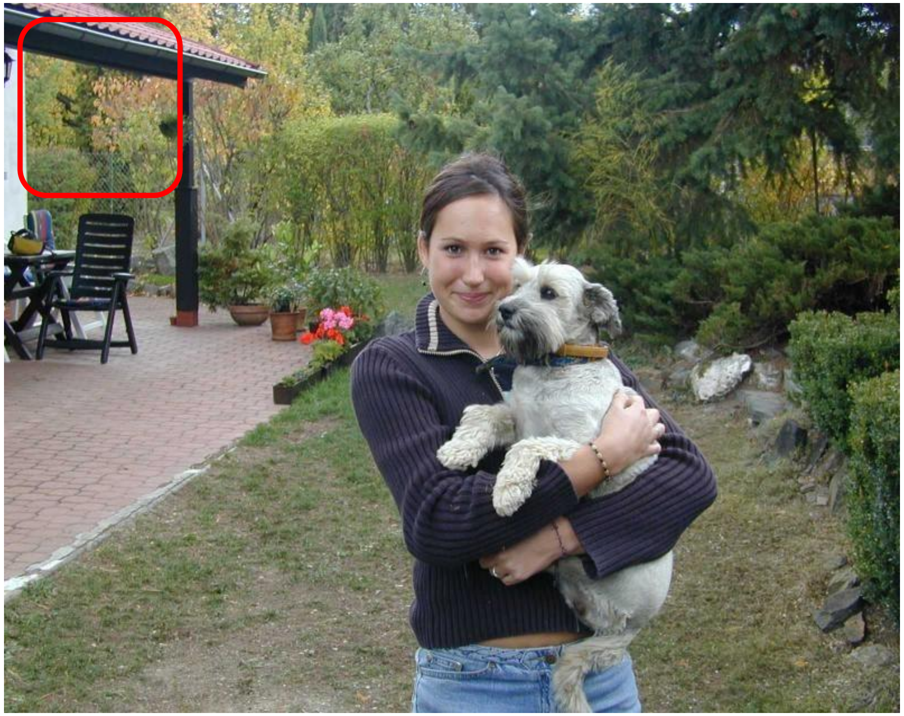
- 1990s: appearance-based models

Slow compute!

- 1990s – present: sliding window approaches

Svetlana Lazebnik

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

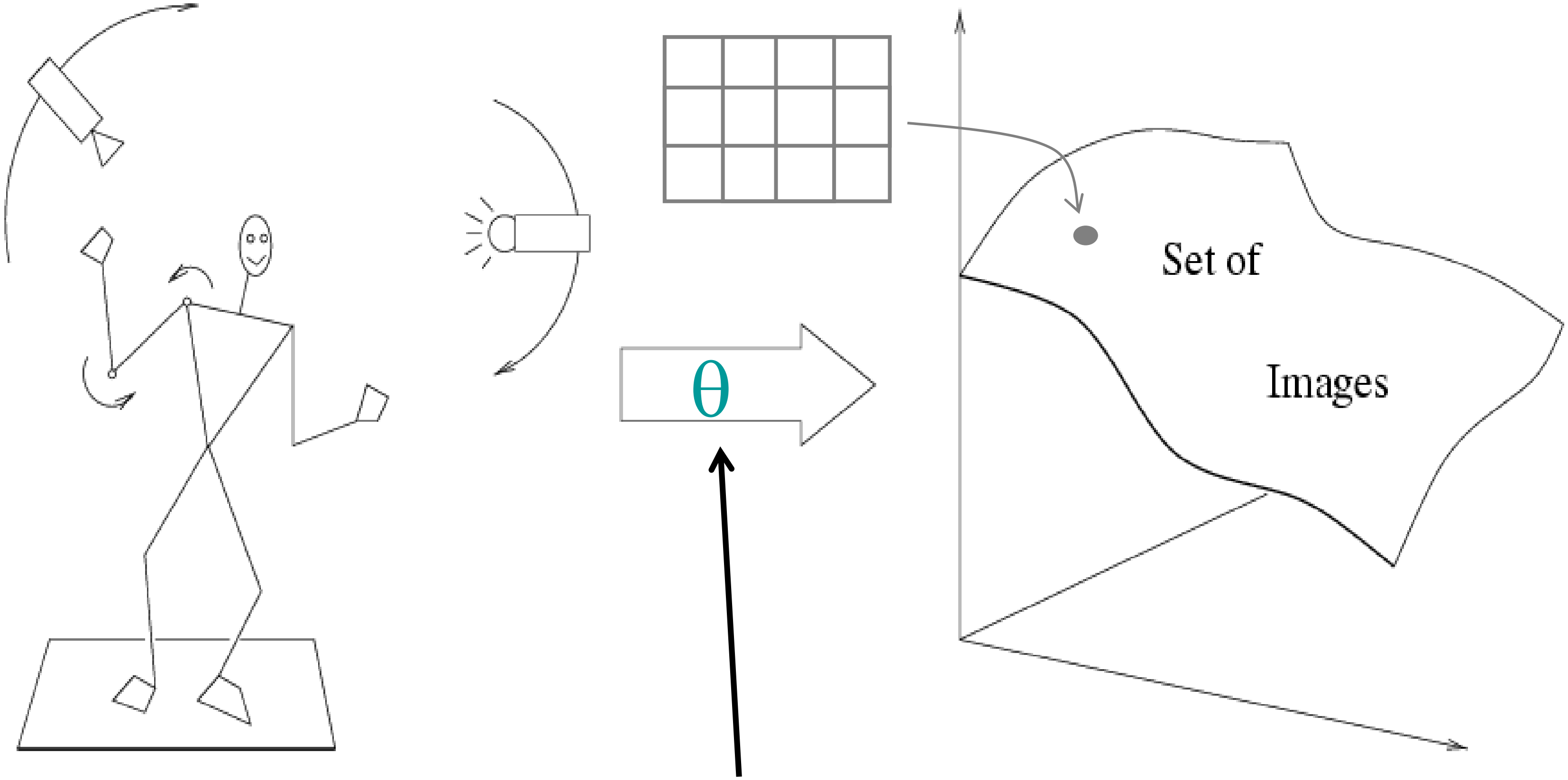
History of ideas in recognition

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

No digital cameras!
Slow compute!

Slow compute!

Svetlana Lazebnik



Variability:

Camera position
Illumination
Shape is partially known

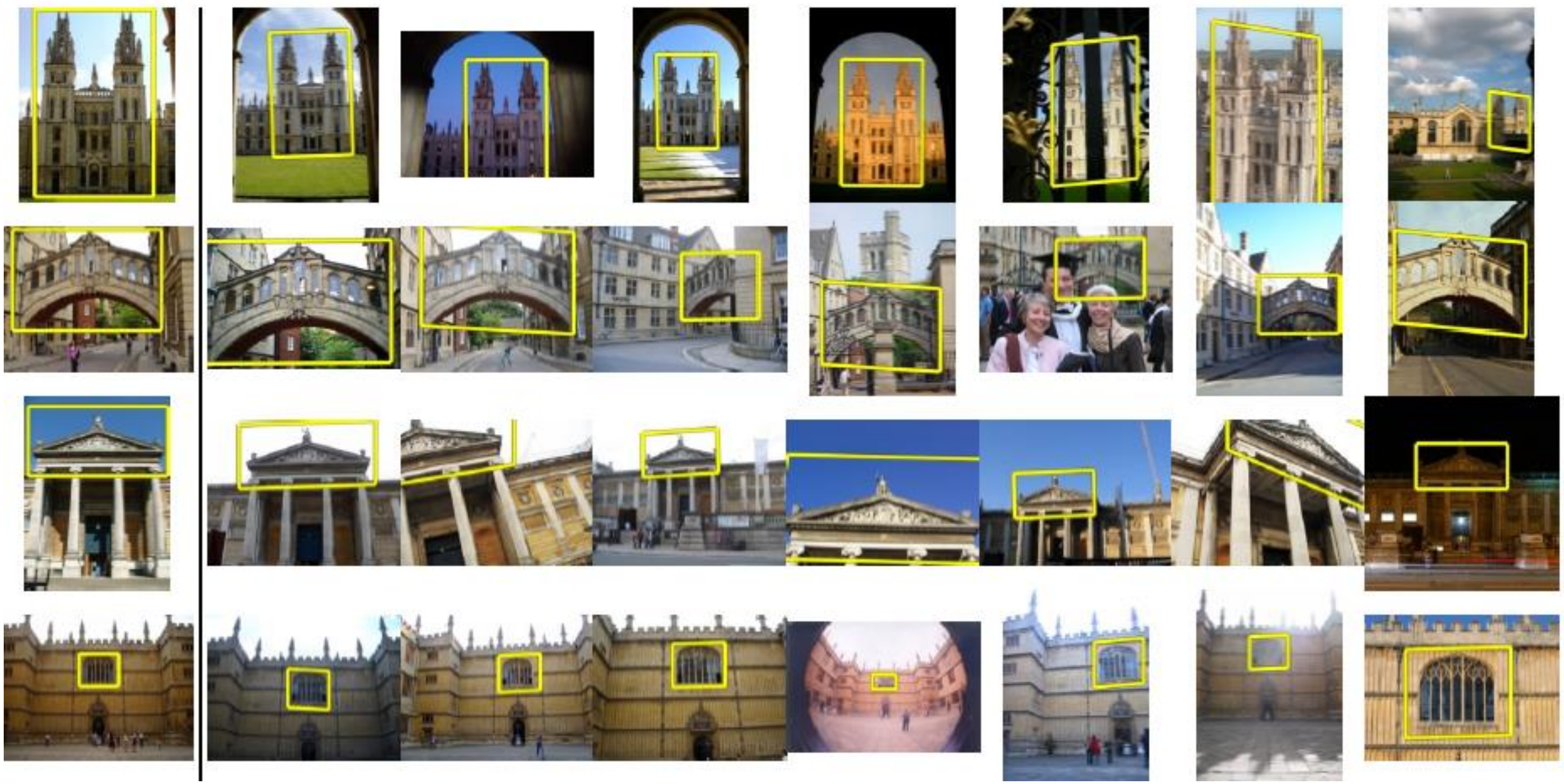
Roberts (1965); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

Local features for object instance recognition

D. Lowe (1999, 2004)

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

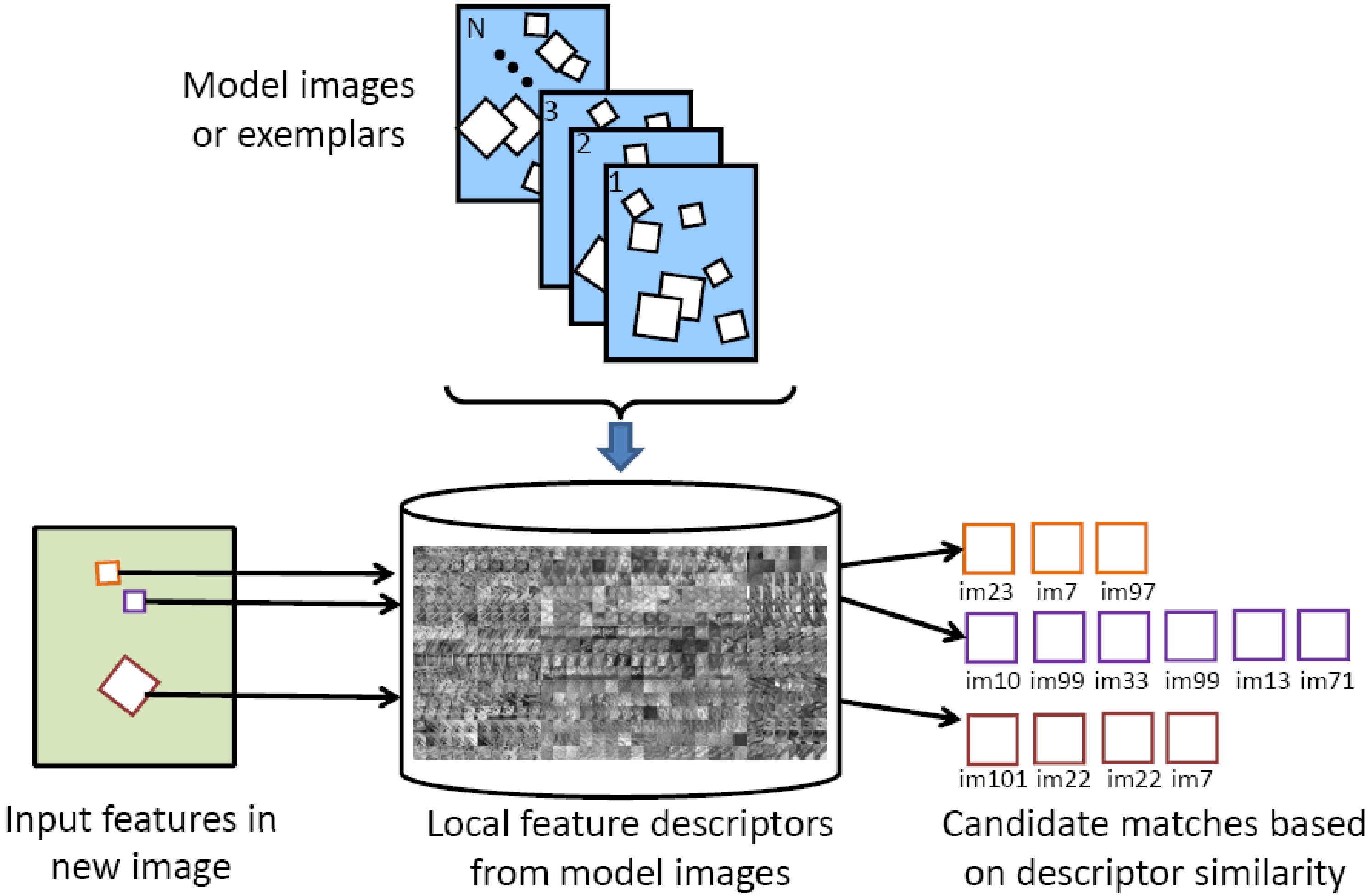


Image credit: K. Grauman and B. Leibe

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

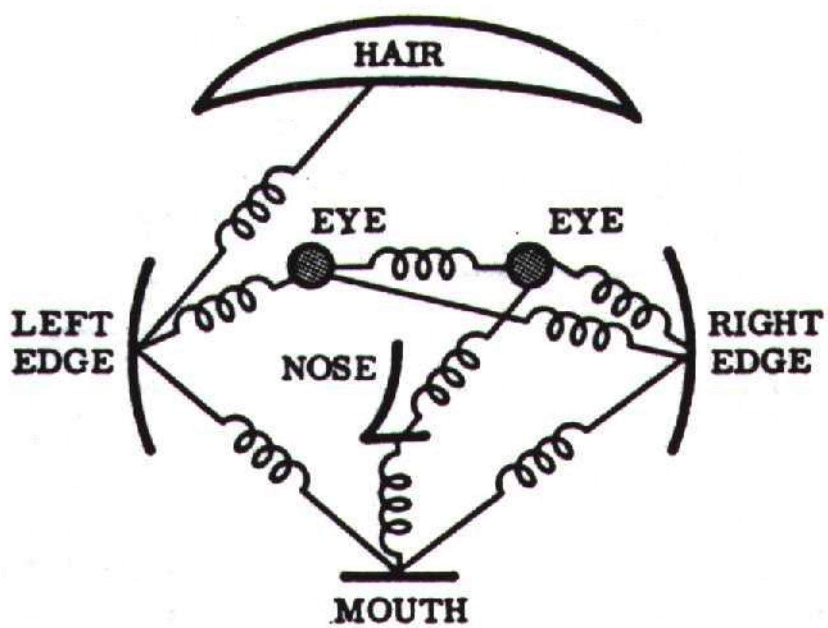


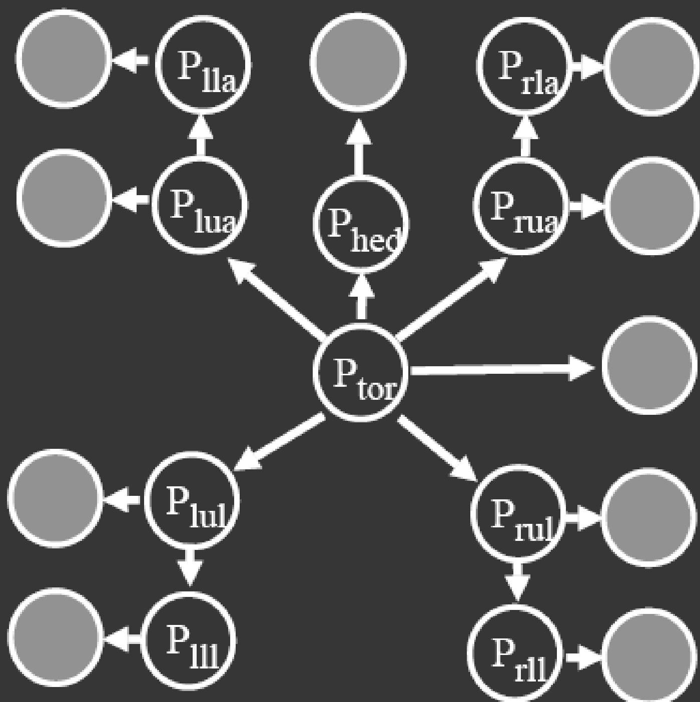
Figure from [Fischler & Elschlager 73]

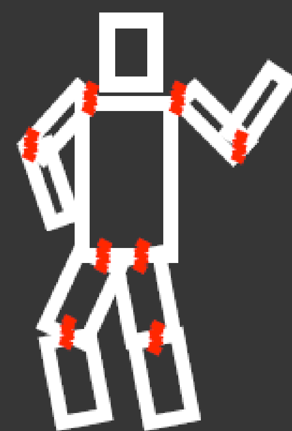
Constellation models

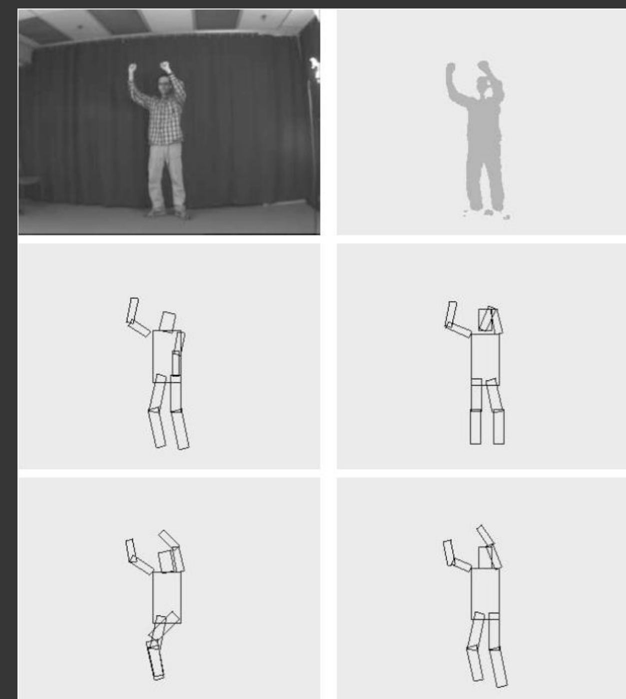
Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

Pictorial structure model

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





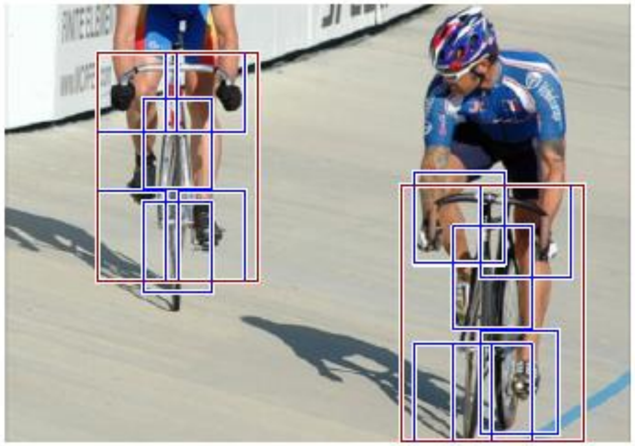


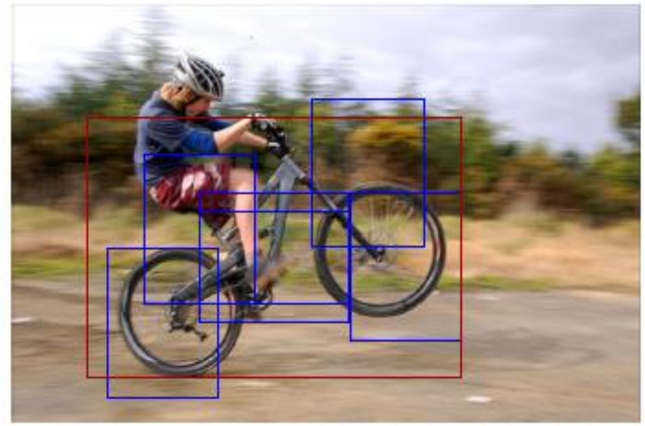
$$\Pr(P_{tor}, P_{arm}, \ldots | \mathrm{Im}) \propto \prod_{i,j} \Pr(P_i \mid P_j) \prod_{i} \Pr(\mathrm{Im}(P_i))$$

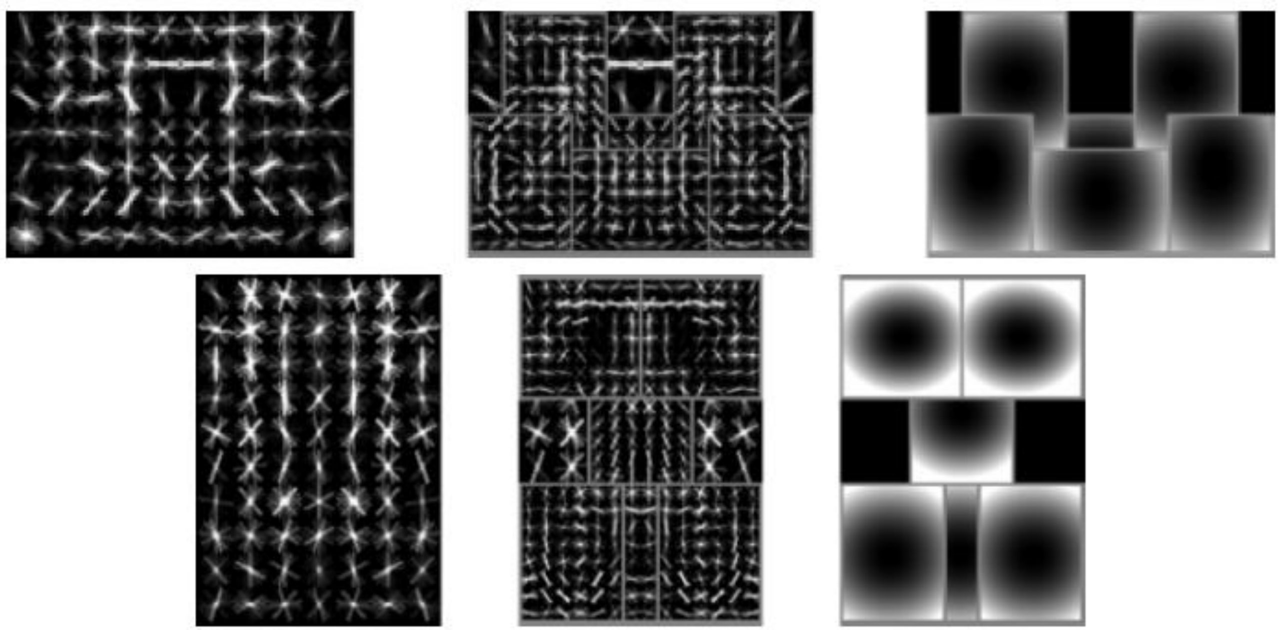
part geometry

part appearance

Discriminatively trained part-based models







P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, PAMI 2009,

"**Object Detection with Discriminatively Trained Part-Based Models**"

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

No digital cameras!
Slow compute!

Slow compute!

Early GPU compute.

Svetlana Lazebnik

Bag-of-features models

Object → **Bag of ‘words’**

Svetlana Lazebnik

Origin 1: Bag-of-words models

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

Origin 1: Bag-of-words models

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

US Presidential Speeches Tag Cloud
`http://chir.ag/phernalia/preztags/`

Origin 1: Bag-of-words models

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

US Presidential Speeches Tag Cloud
`http://chir.ag/phernalia/preztags/`

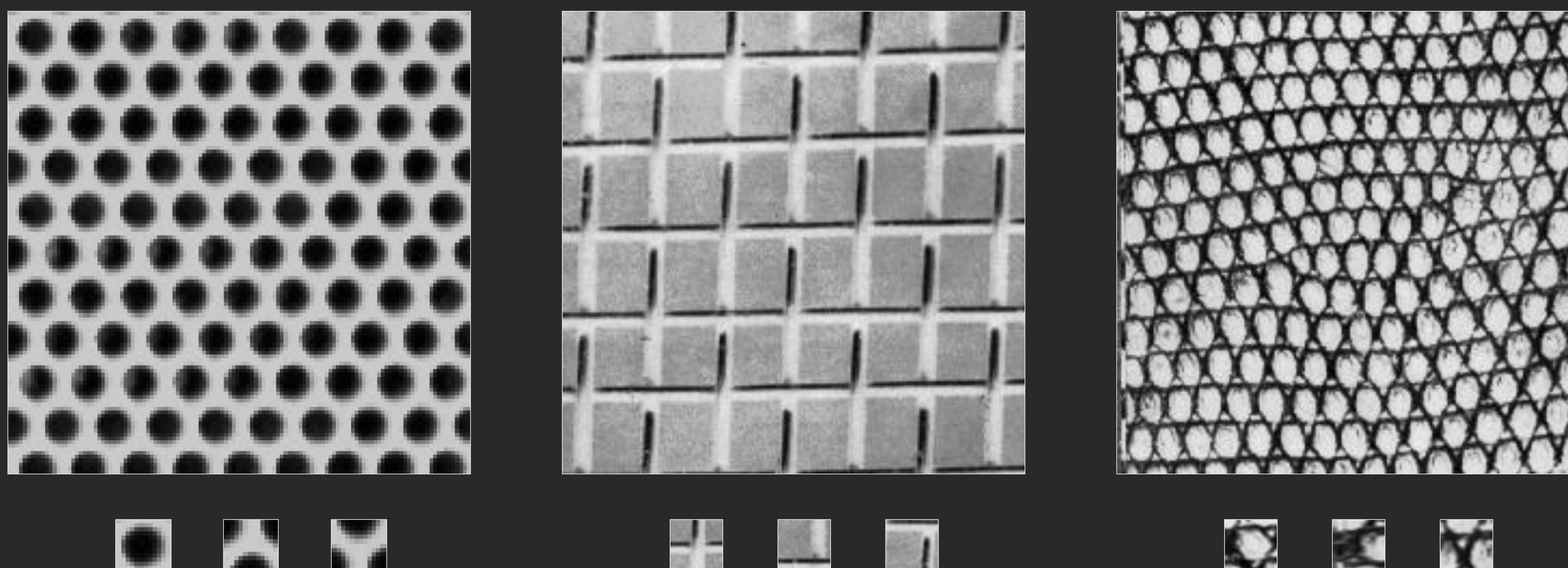
Origin 1: Bag-of-words models

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

US Presidential Speeches Tag Cloud
`http://chir.ag/phernalia/preztags/`

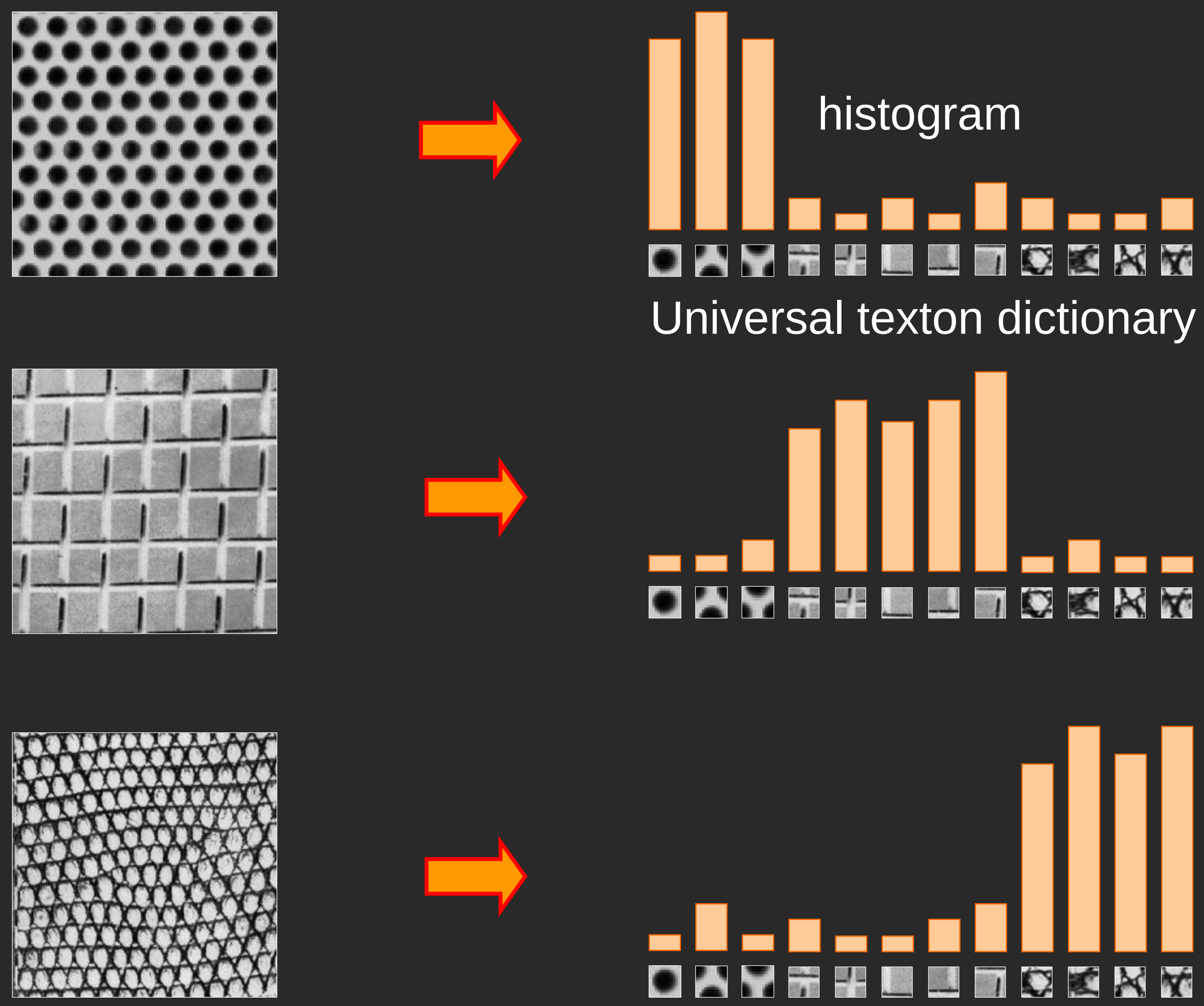
Origin 2: Texture recognition

- Characterized by repetition of basic elements or *textons*
- For stochastic textures, the identity of textons matters, not their spatial arrangement



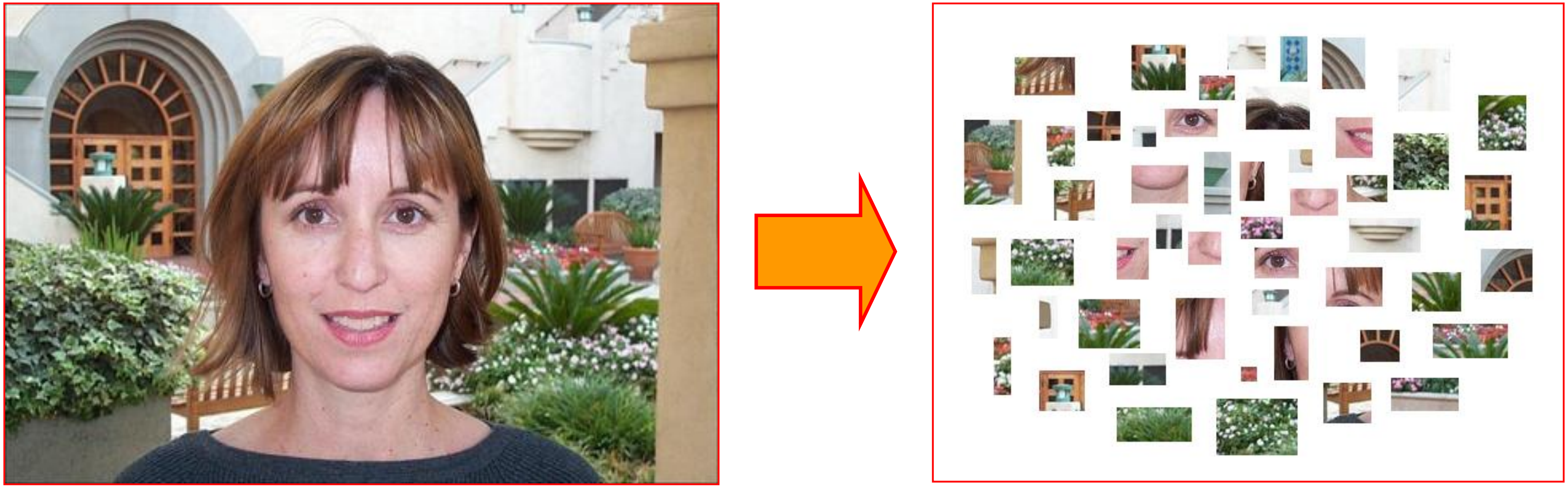
Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Origin 2: Texture recognition



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

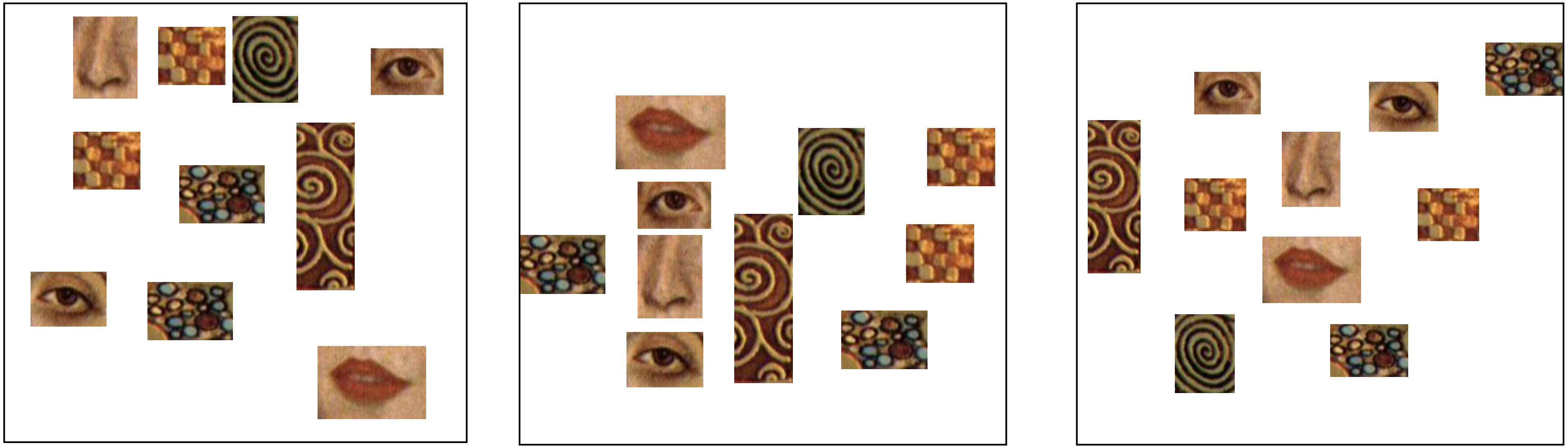
Bag-of-features models



Svetlana Lazebnik

Objects as texture

- All of these are treated as being the same

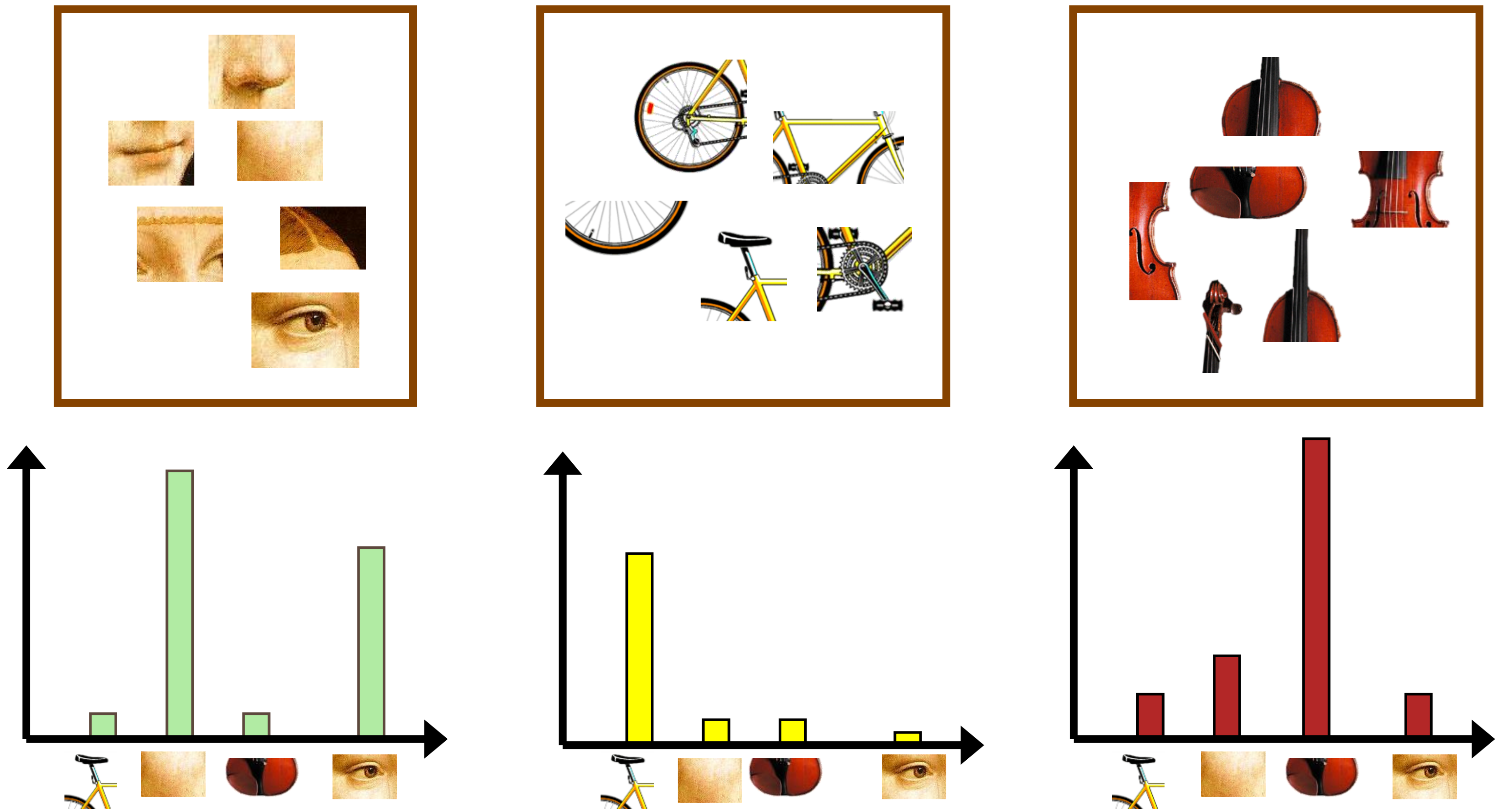


- No distinction between foreground and background: scene recognition?

Svetlana Lazebnik

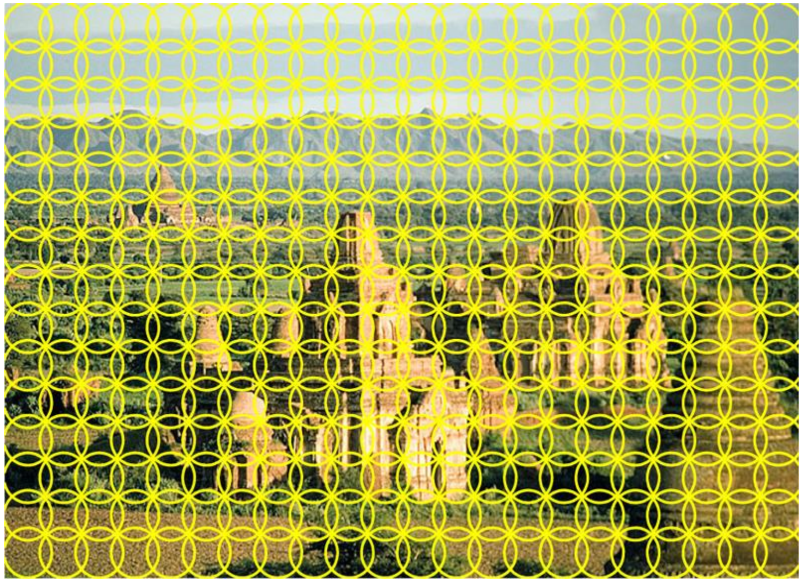
Bag-of-features steps

1. Feature extraction
2. Learn "visual vocabulary"
3. Quantize features using visual vocabulary
4. Represent images by frequencies of "visual words"

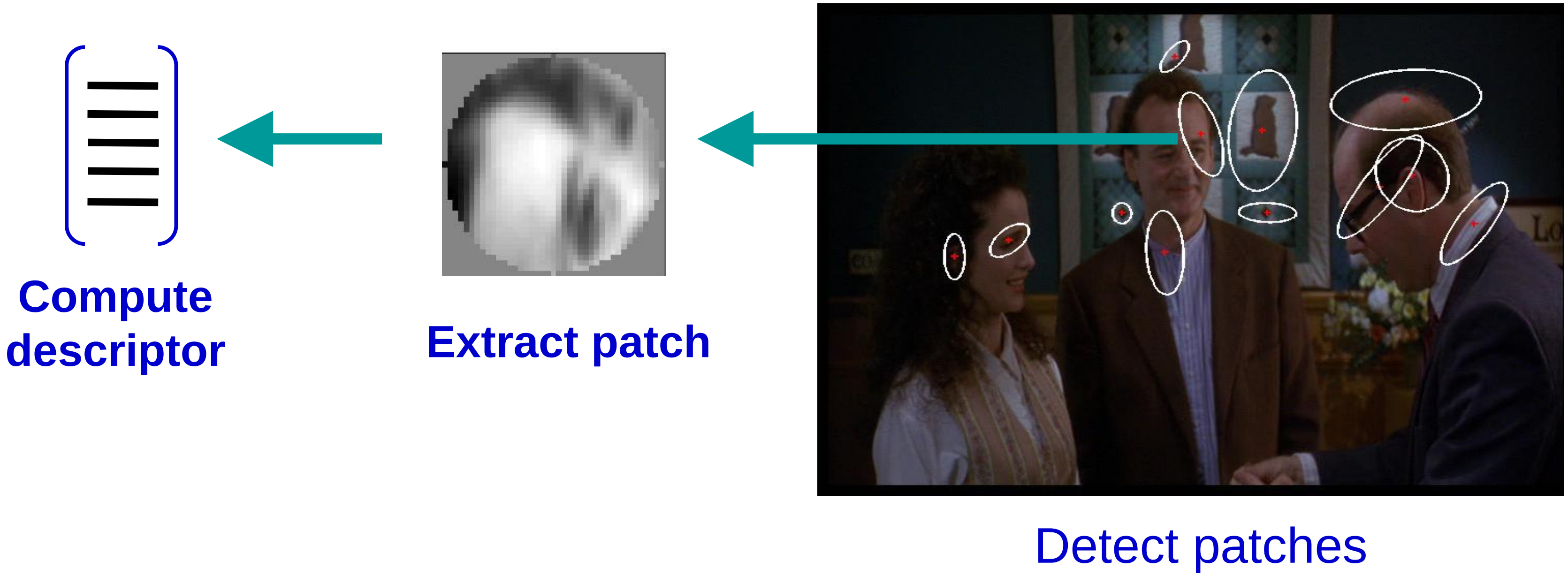


1. Feature extraction

- Regular grid or interest regions

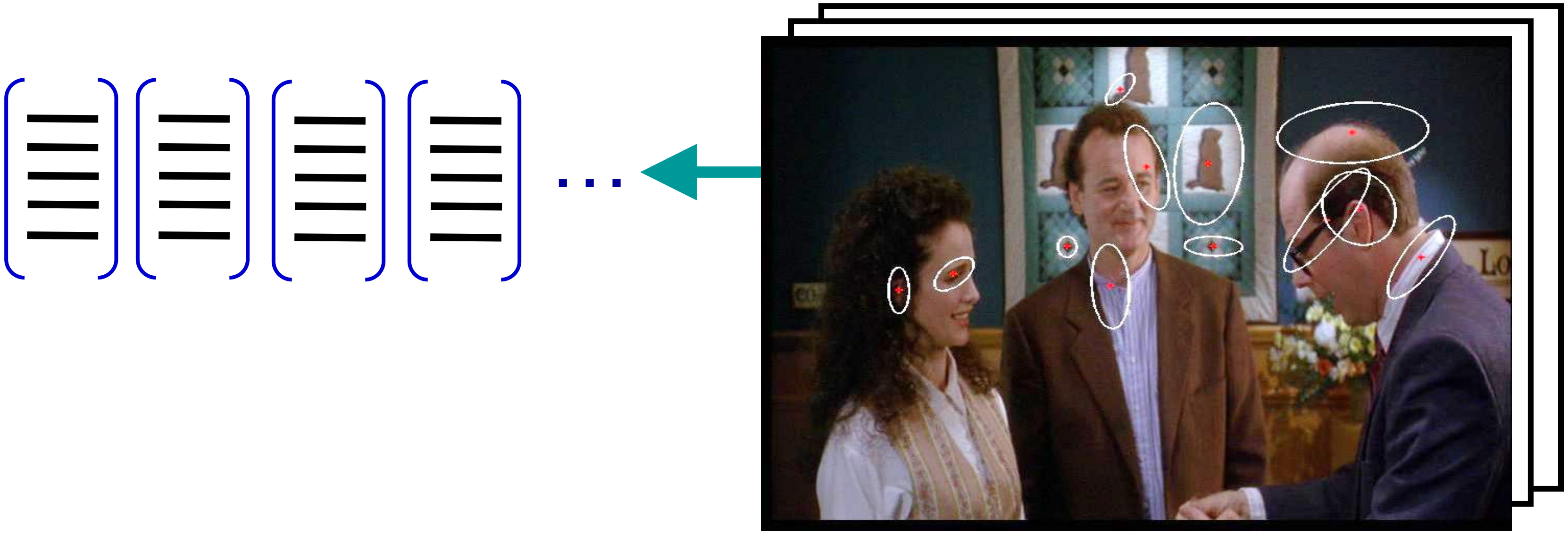


1. Feature extraction



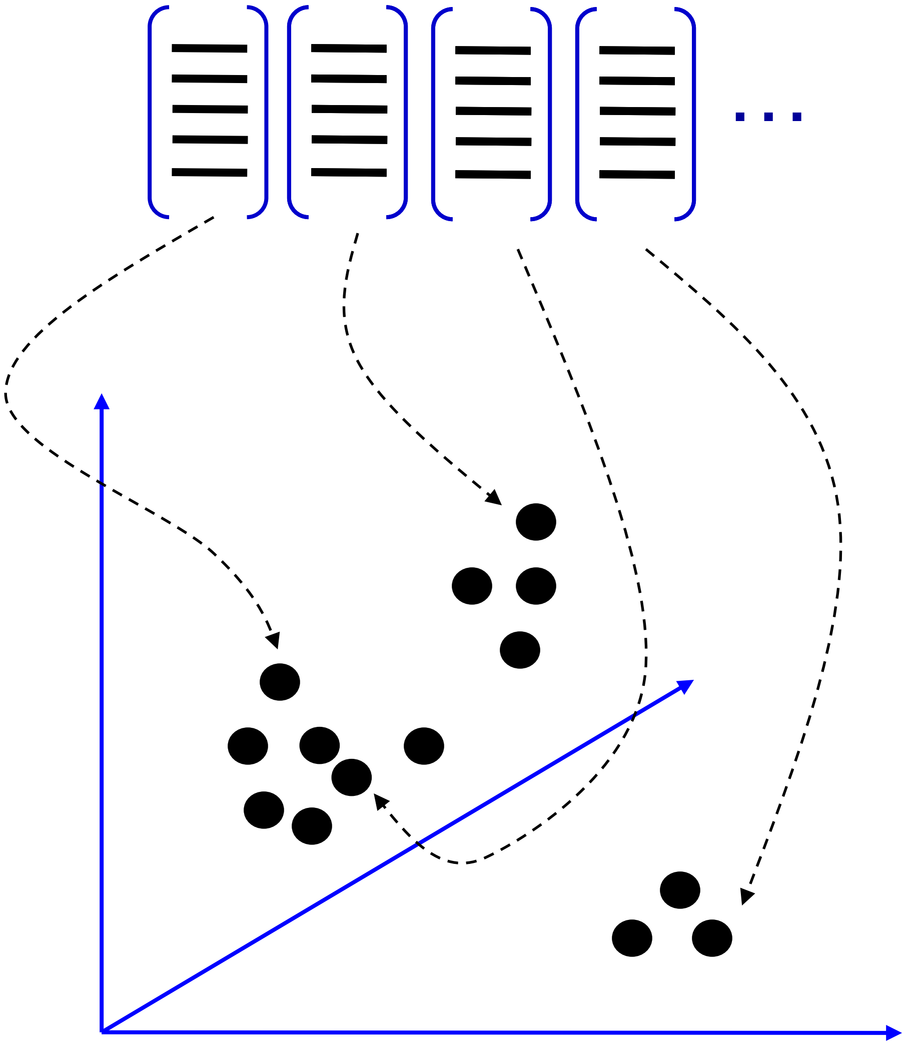
Slide credit: Josef Sivic

1. Feature extraction



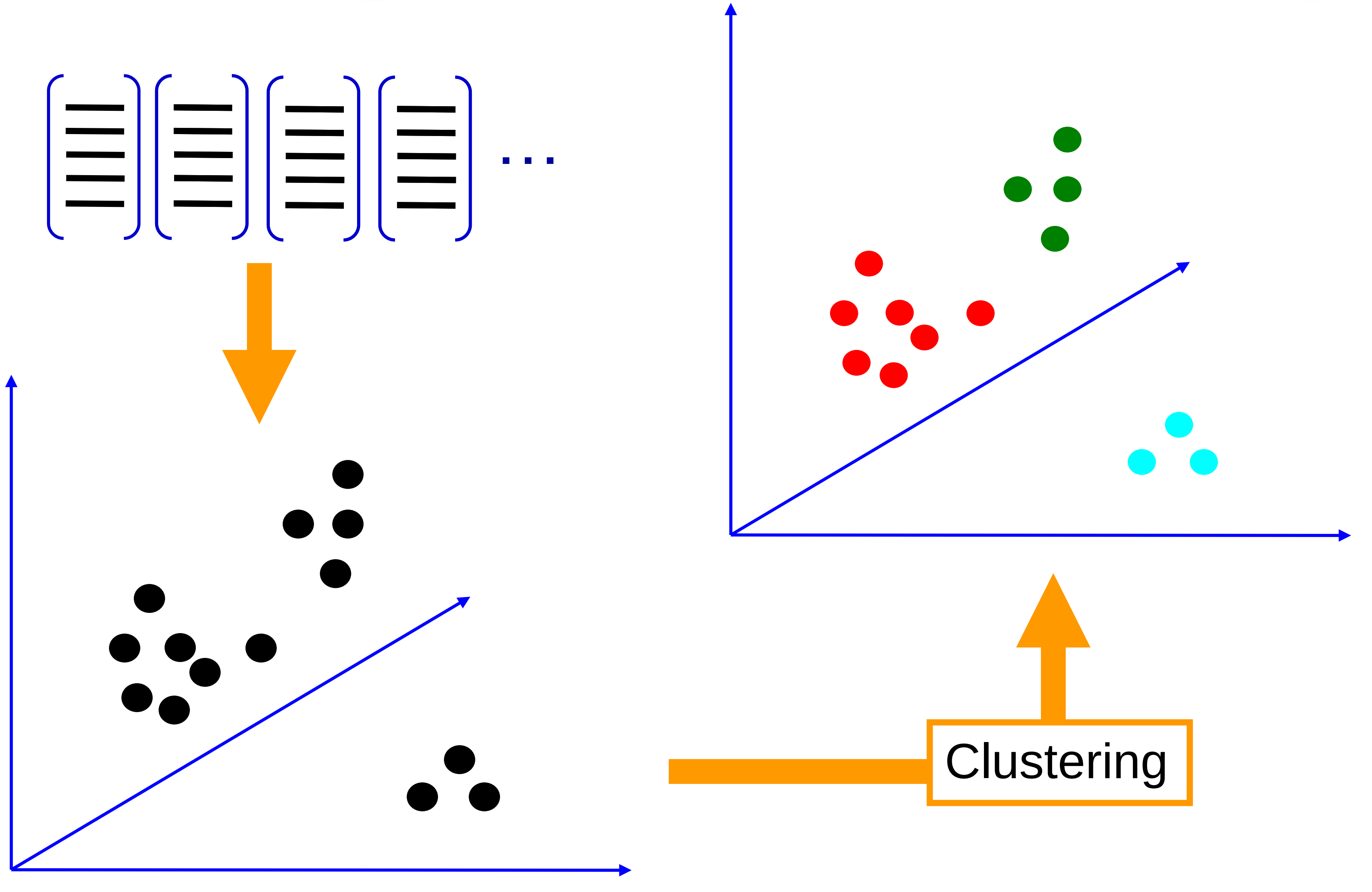
Slide credit: Josef Sivic

2. Learning the visual vocabulary



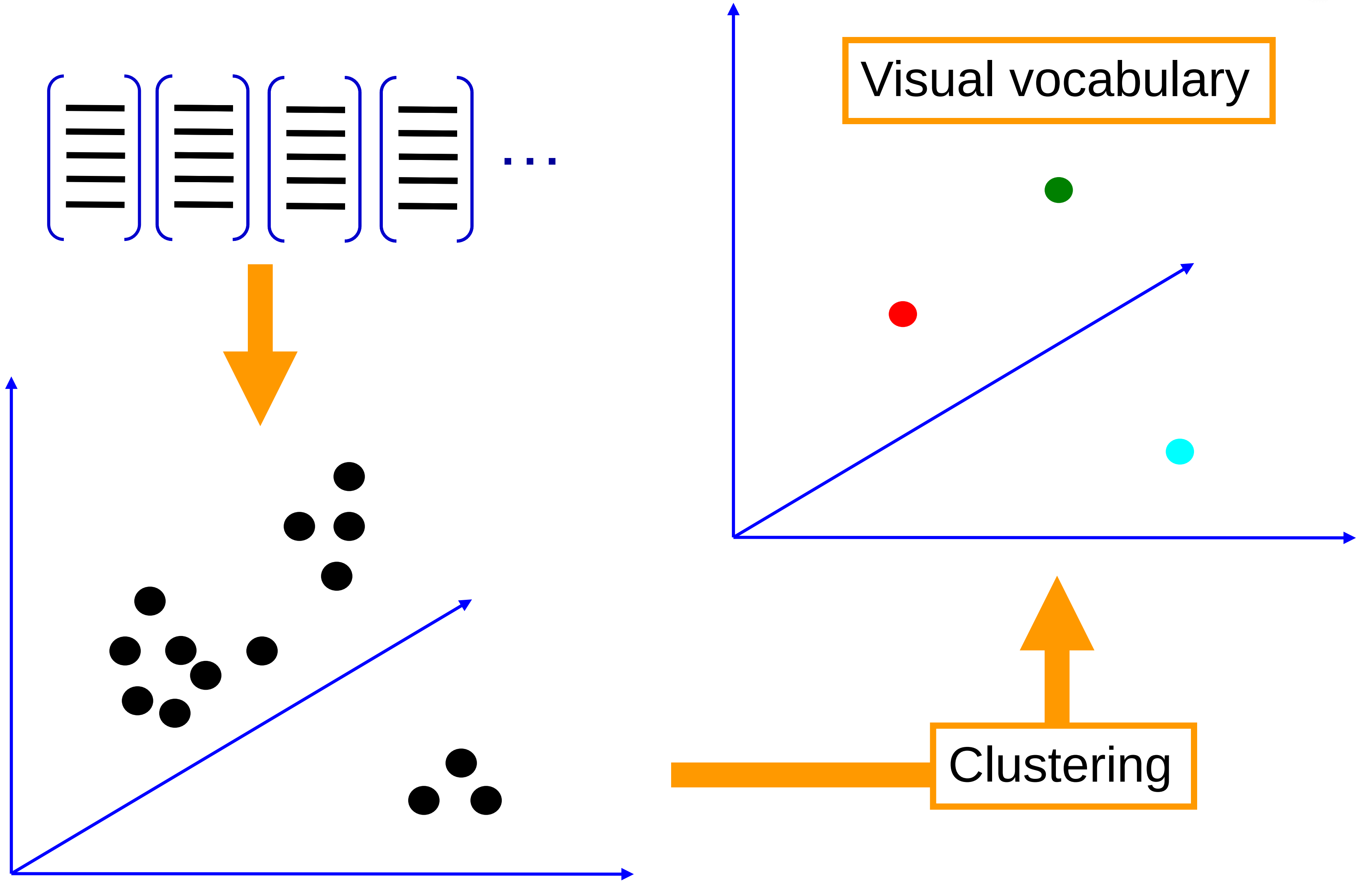
Slide credit: Josef Sivic

2. Learning the visual vocabulary



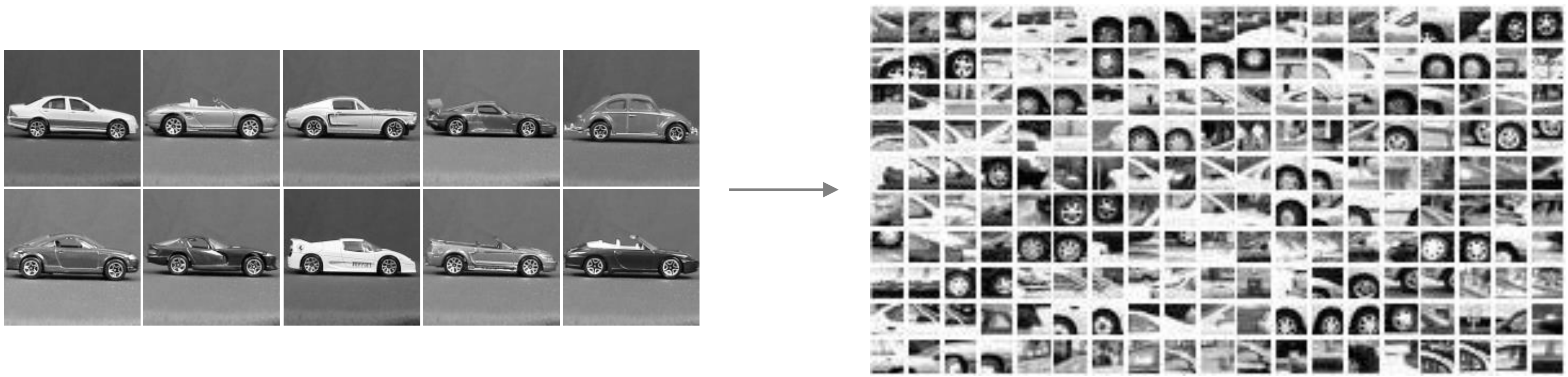
Slide credit: Josef Sivic

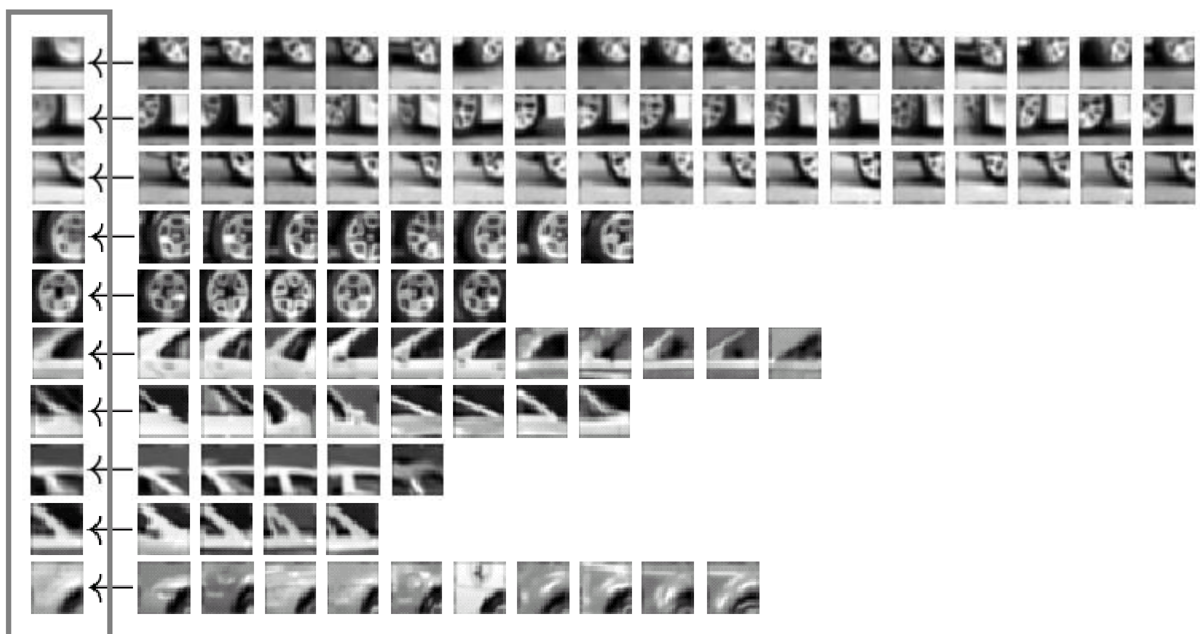
3. Quantize the visual vocabulary



Slide credit: Josef Sivic

Example codebook





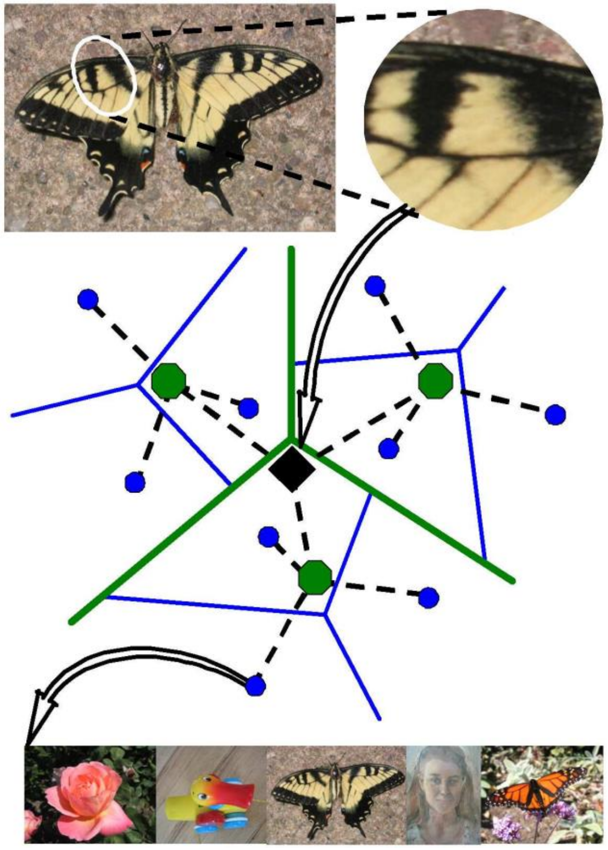
...

Appearance codebook

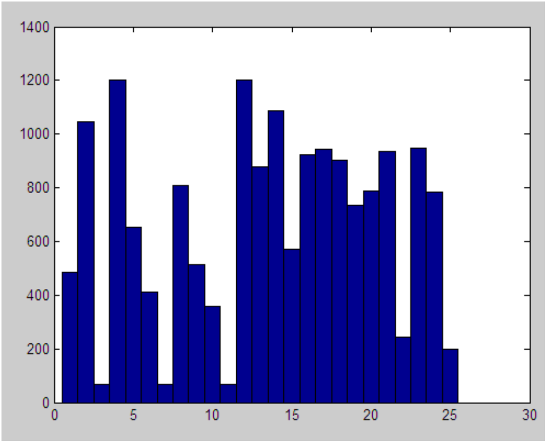
Source: B. Leibe

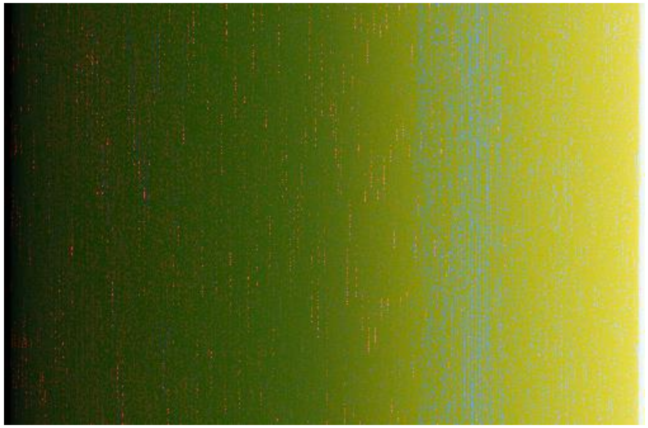
Visual vocabularies: Issues

- 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)



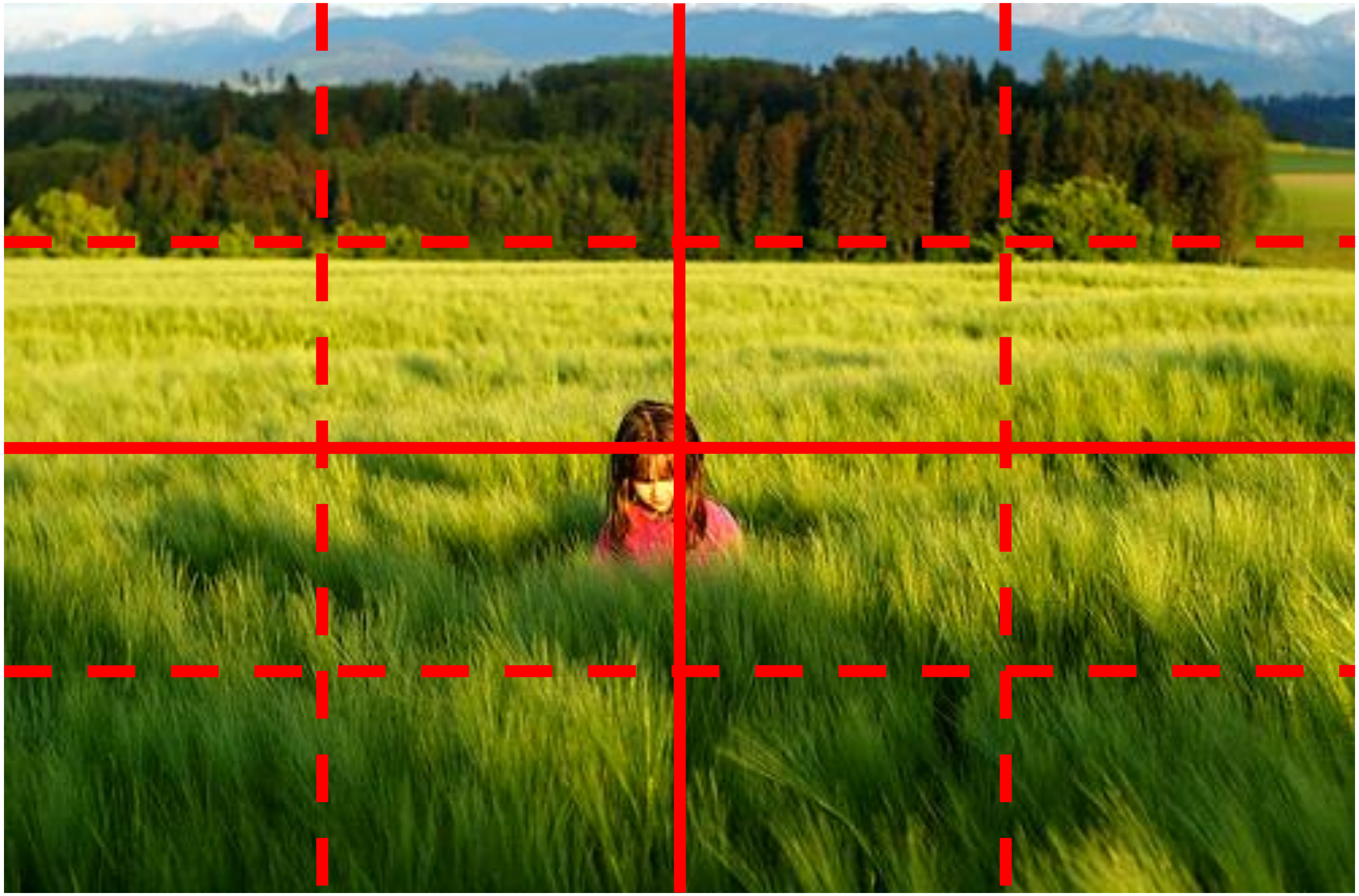
But what about layout?





All of these images have the same color histogram

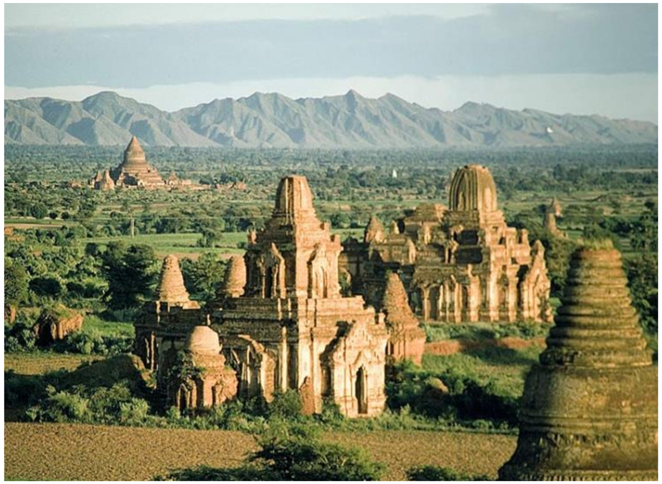
Spatial pyramid

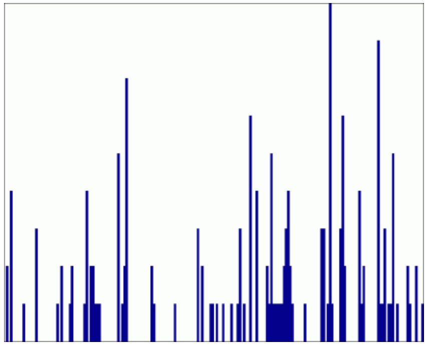


Compute histogram in each spatial bin

Spatial pyramid representation

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



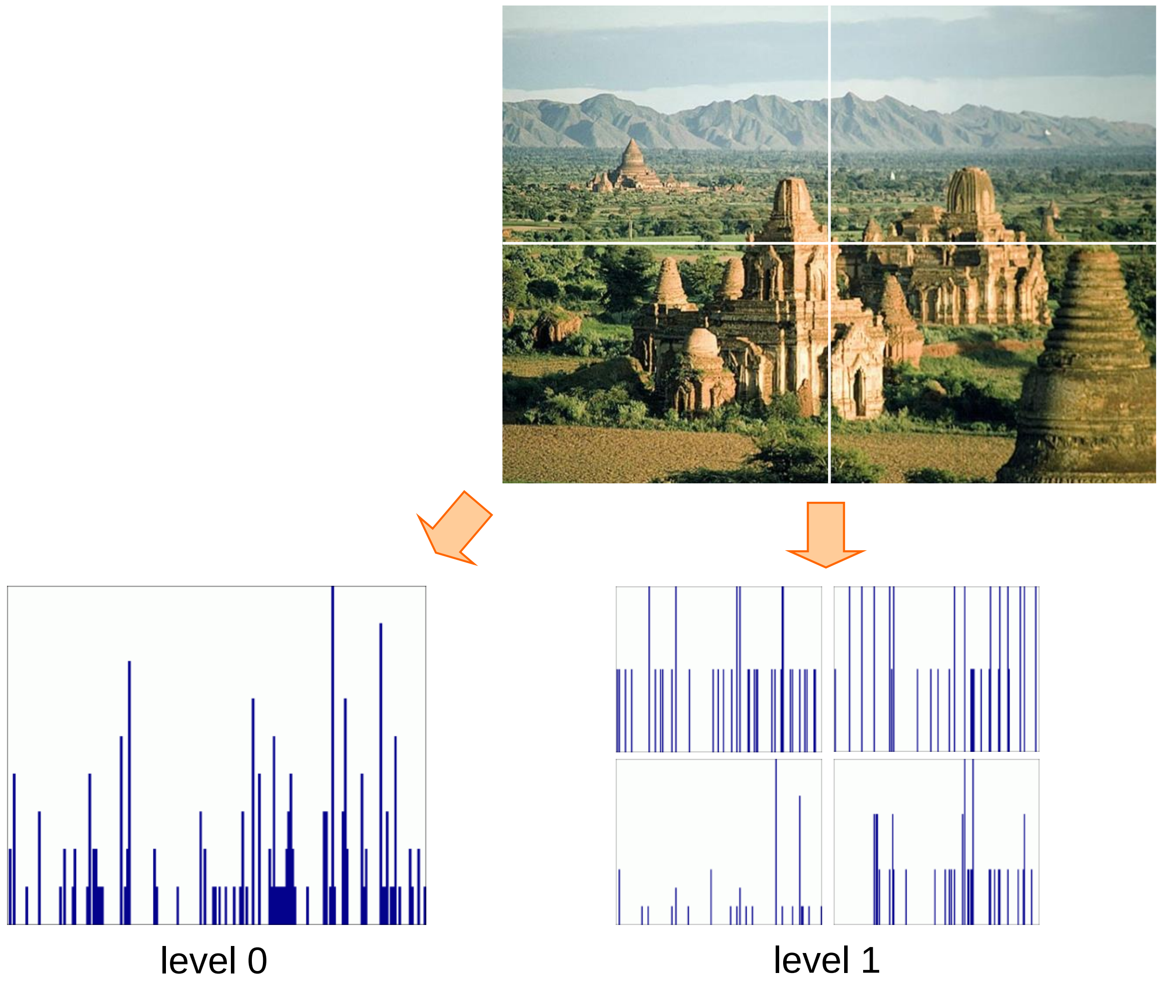


level 0

Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

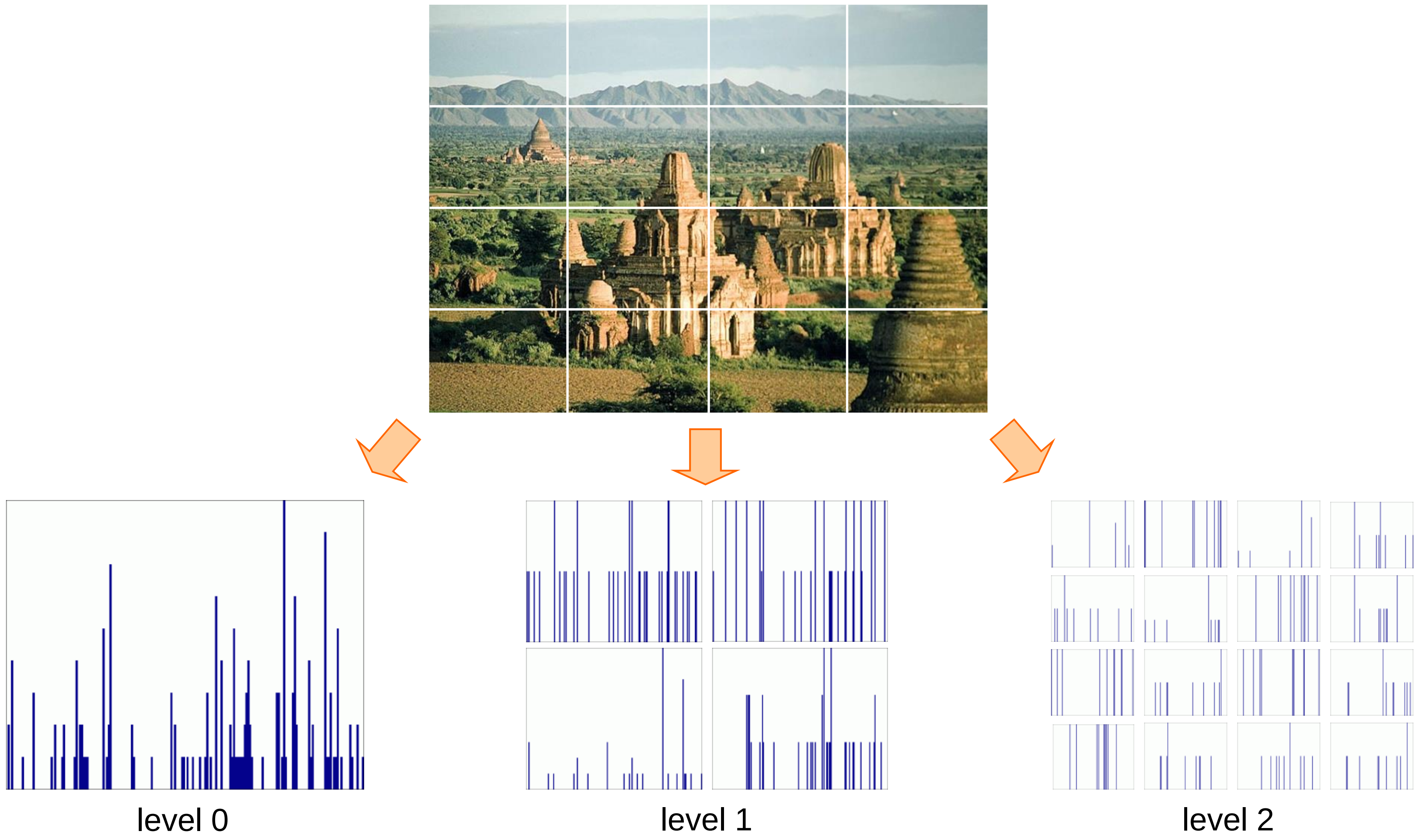
- Extension of a bag of features
- Locally orderless representation at several levels of resolution



Lazebnik, Schmid & Ponce (CVPR 2006)

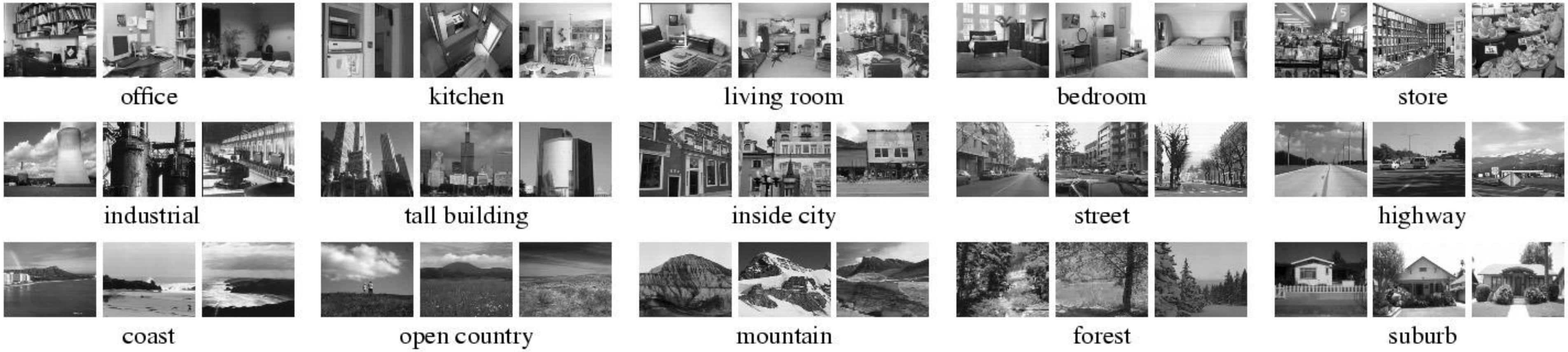
Spatial pyramid representation

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



Lazebnik, Schmid & Ponce (CVPR 2006)

Scene category dataset

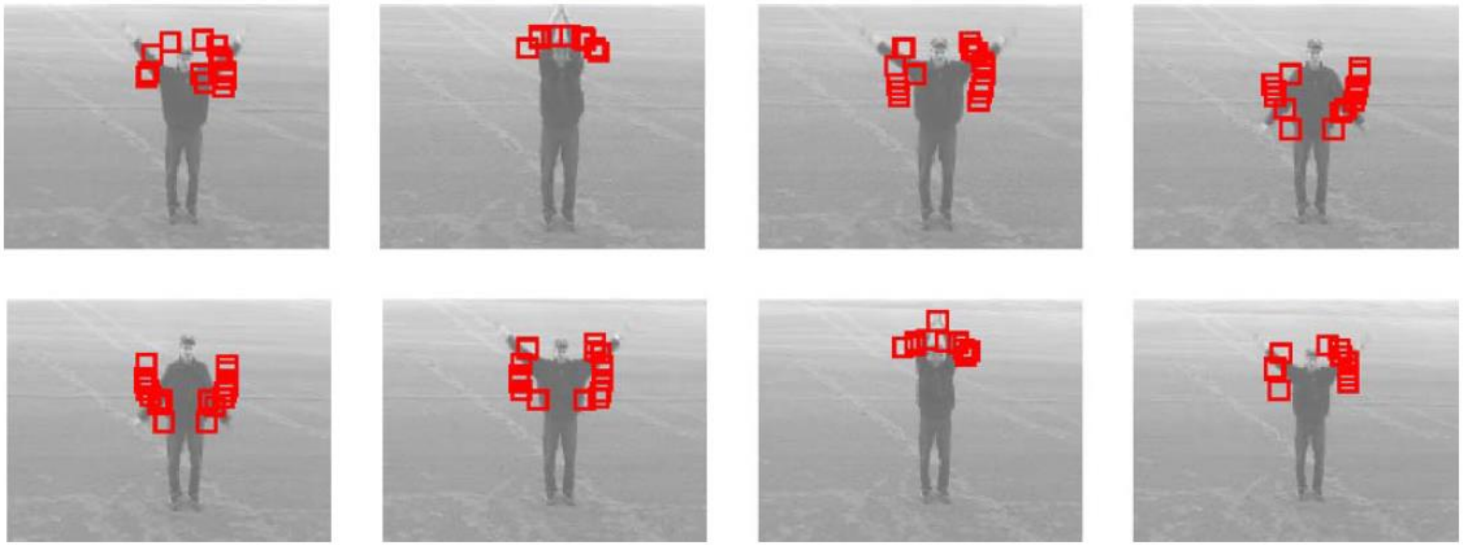


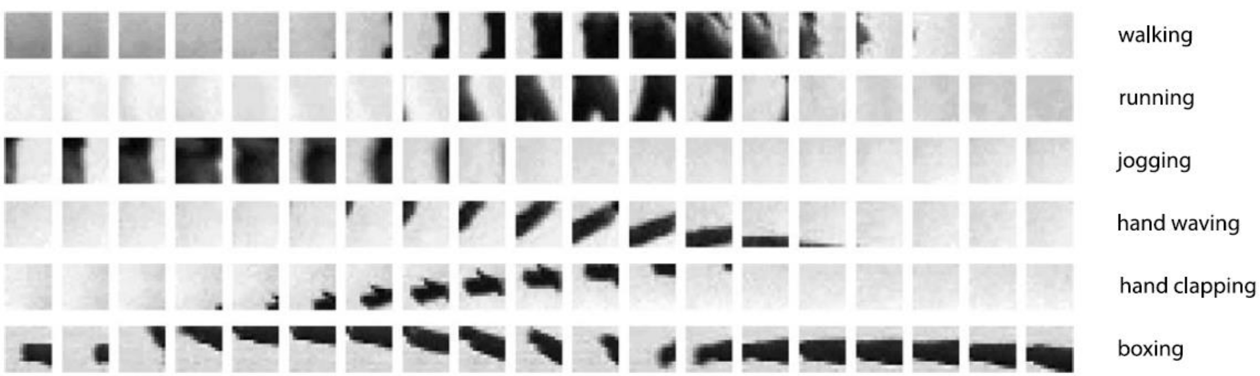
Multi-class classification results
(100 training images per class)

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

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 — No digital cameras! Slow compute!
- 1990s: appearance-based models — Slow compute!
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features — Early GPU compute.
- *Present trends:*
  Combined local and global methods, context, deep learning — GPU/cloud compute.

Svetlana Lazebnik