

Machine Learning Problems

Supervised Learning

Unsupervised Learning

Discrete
Continuous

classification or
categorization

clustering

regression

dimensionality
reduction

Supervised learning

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

The diagram shows the equation $y = f(\mathbf{x})$ in blue. Below the equation, three red arrows point upwards to the components: the first arrow points to y and is labeled 'output'; the second arrow points to f and is labeled 'prediction function'; the third arrow points to \mathbf{x} and is labeled 'Image feature'.

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* \mathbf{x} and output the predicted value $y = f(\mathbf{x})$

Image Categorization

Training

Training Labels

Training Images

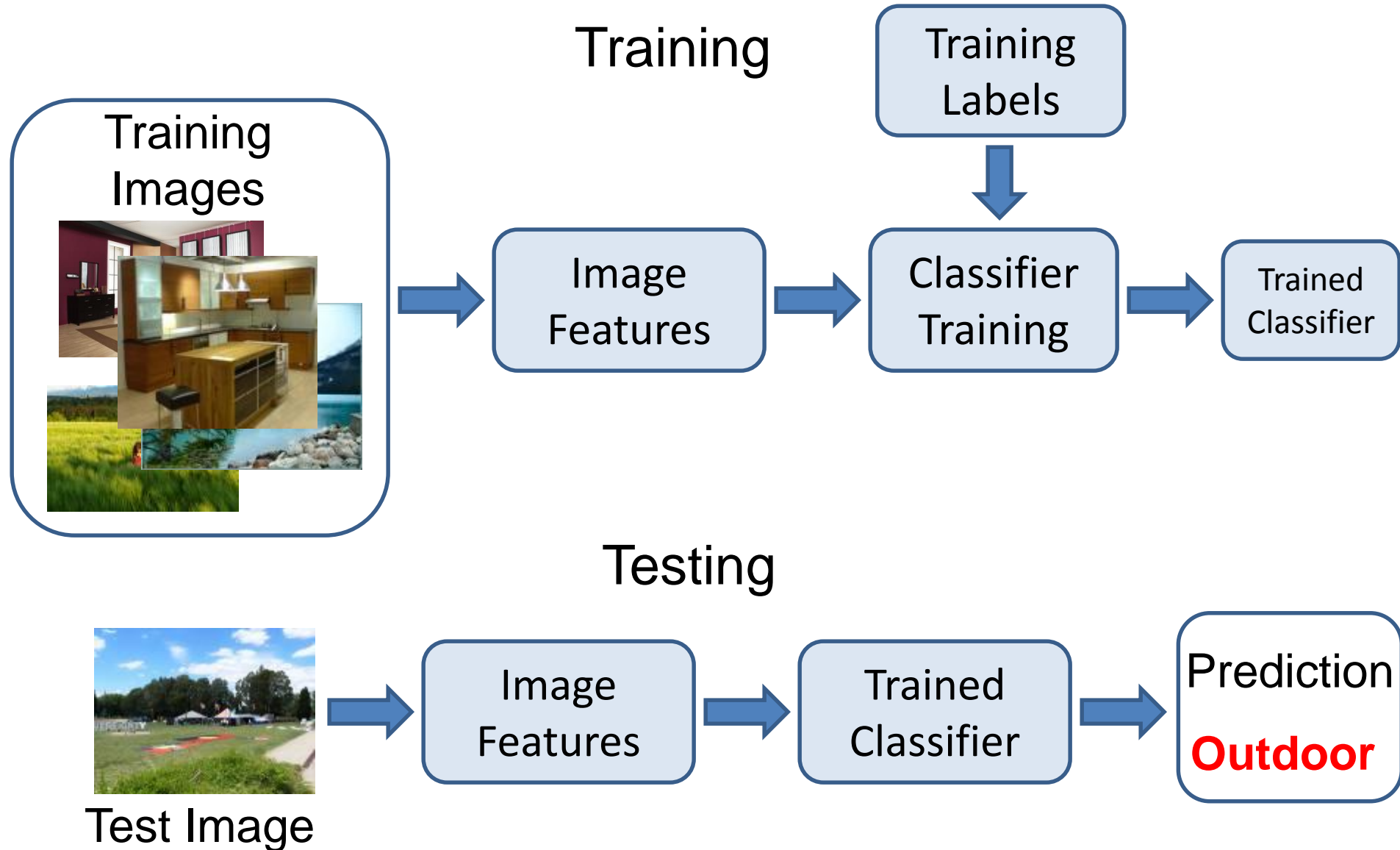
Image Features

Classifier Training

Trained Classifier



Image Categorization

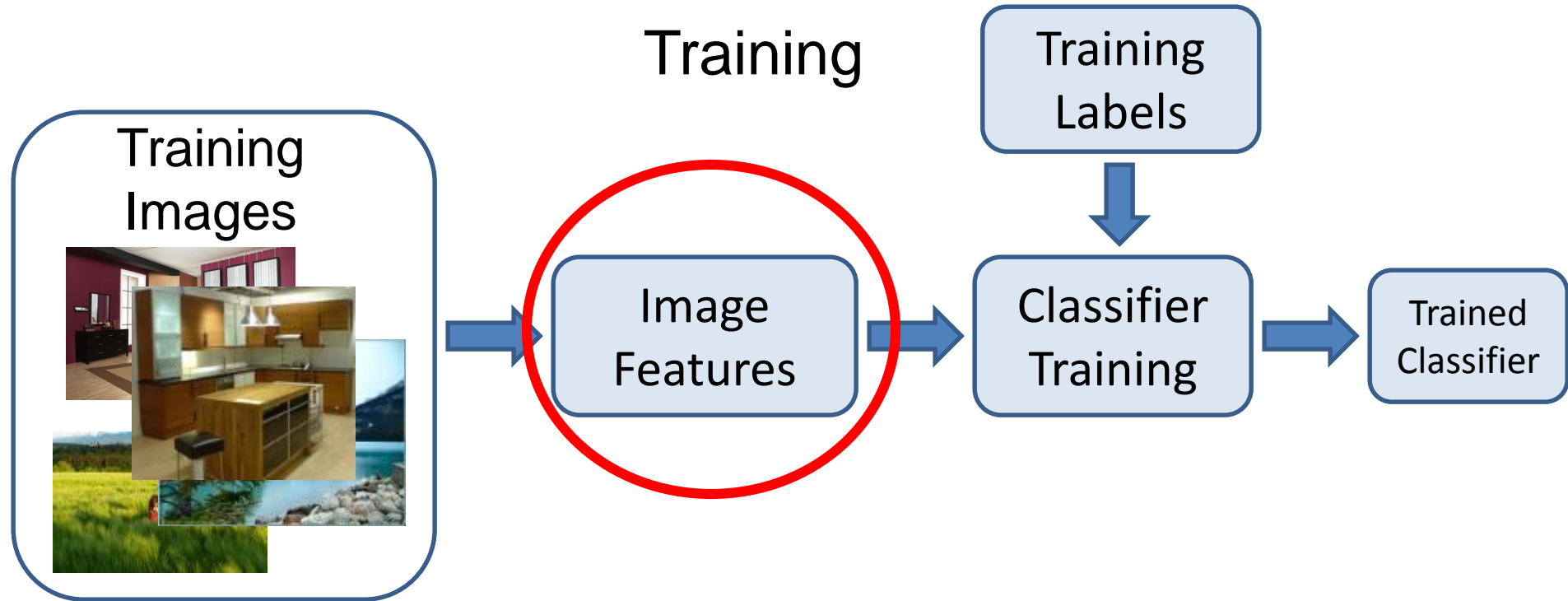


Example: Scene Categorization

- Is this a kitchen?

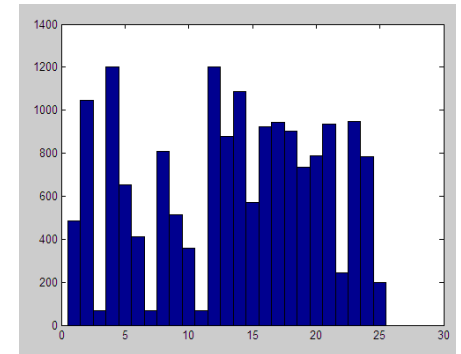


Image features



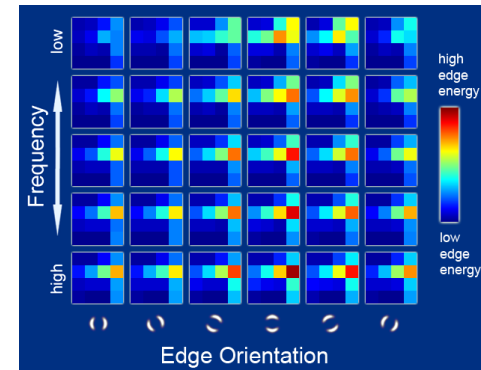
Features

- Raw pixels



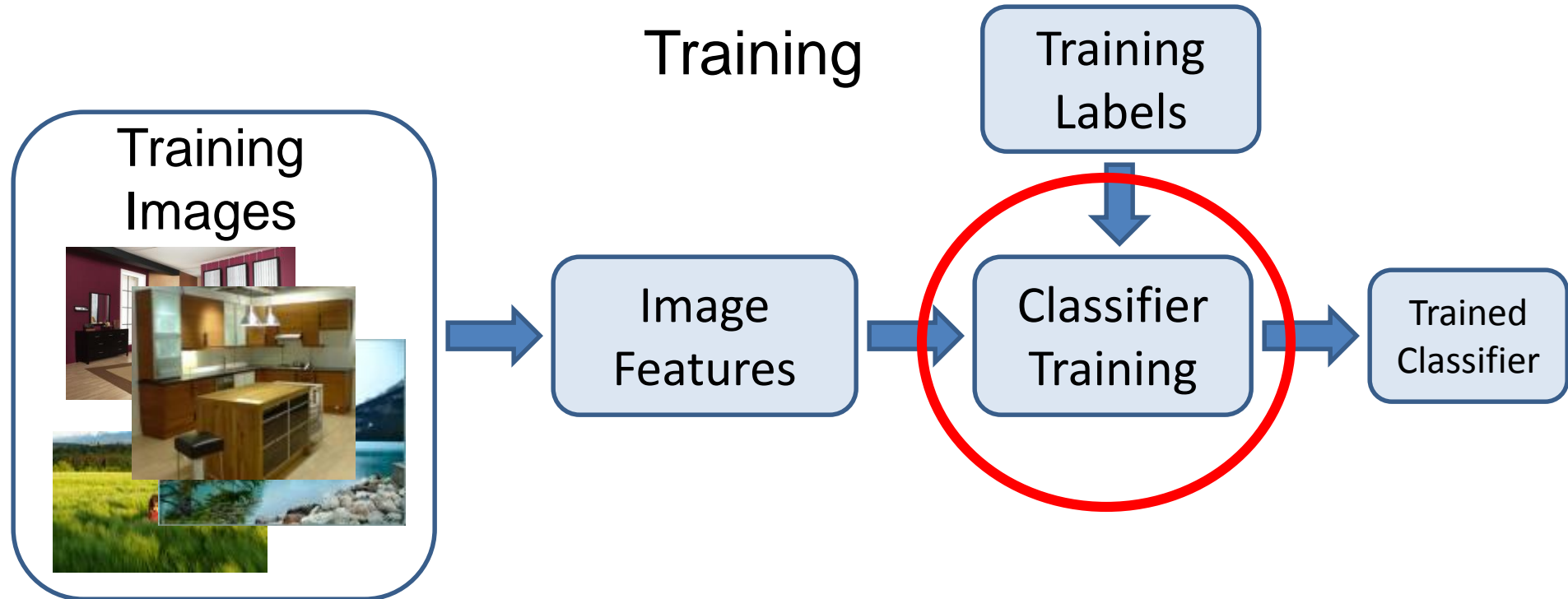
- Histograms

- GIST descriptors



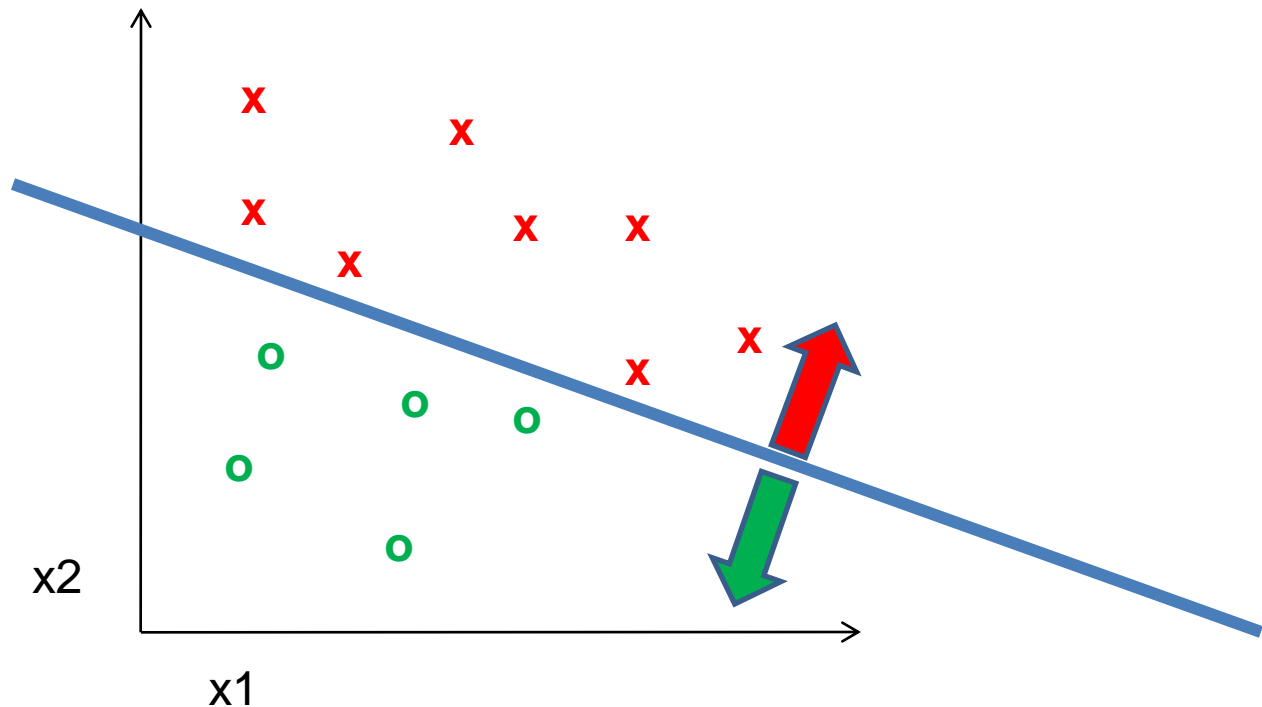
- ...

Classifiers



Learning a classifier

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

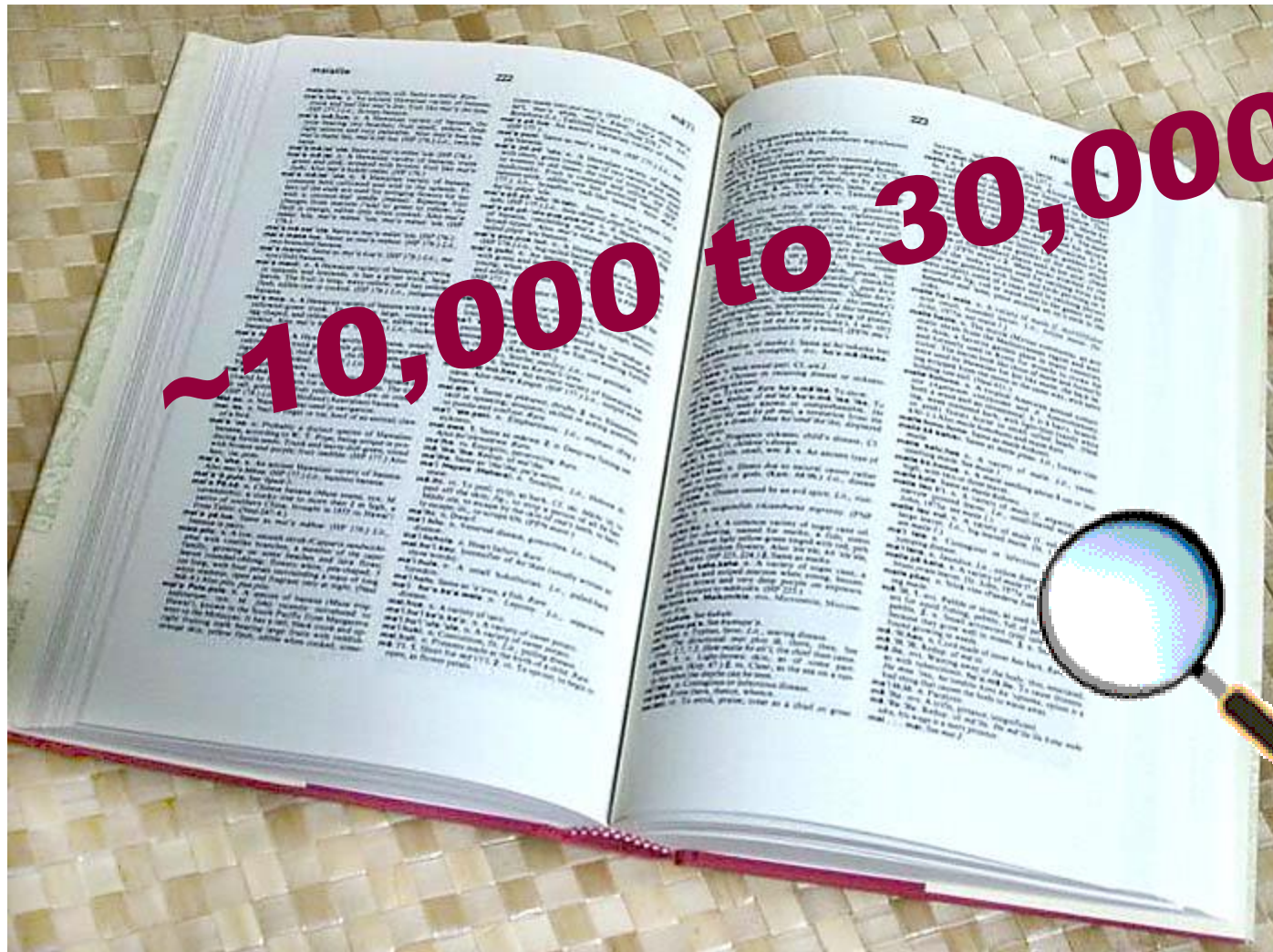


Recognition: Overview and History



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

How many visual object categories are there?

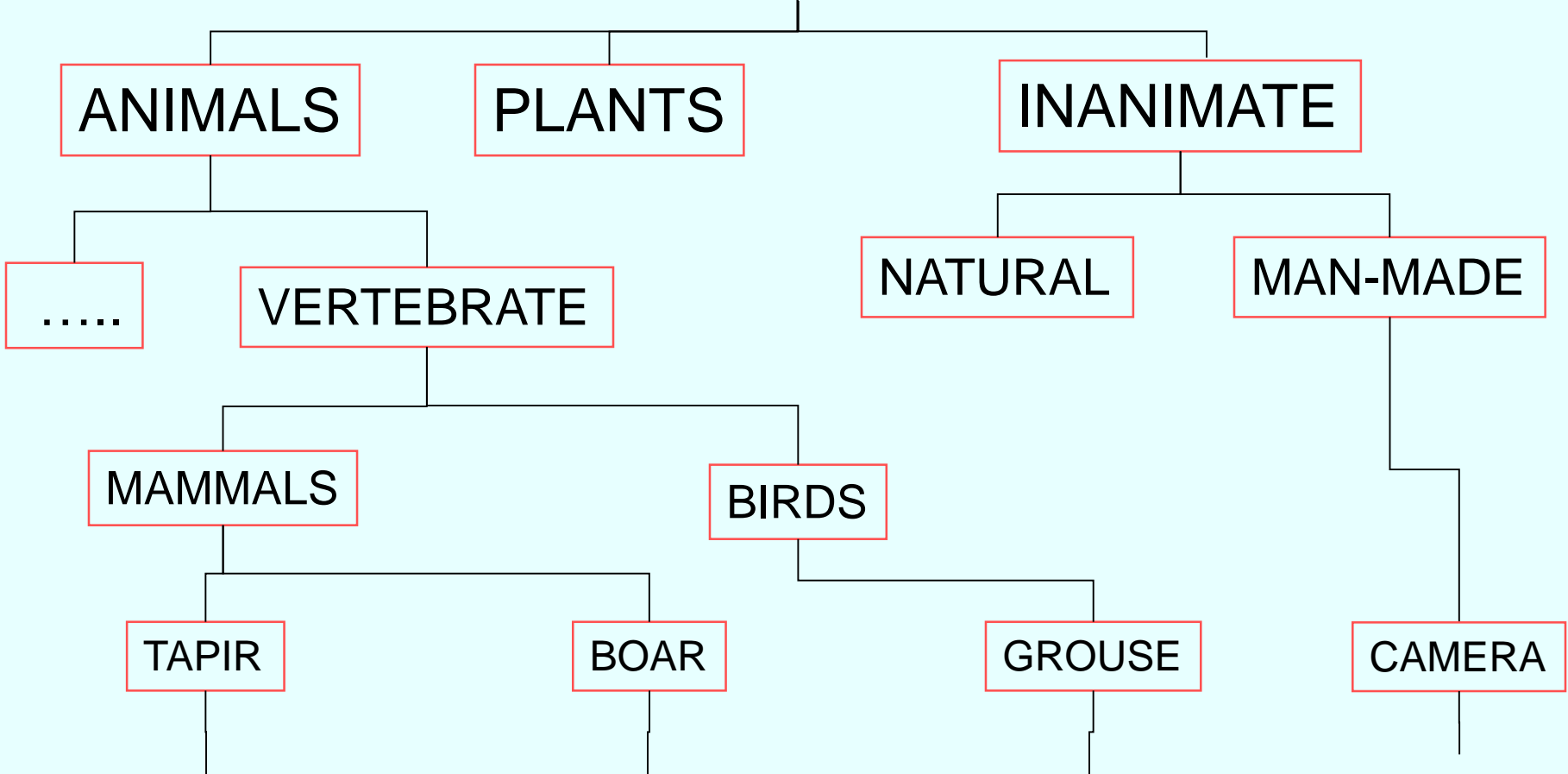




~10,000 to 30,000



OBJECTS



Specific recognition tasks



Scene categorization or classification



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

Image annotation / tagging / attributes



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

Object detection

- find pedestrians

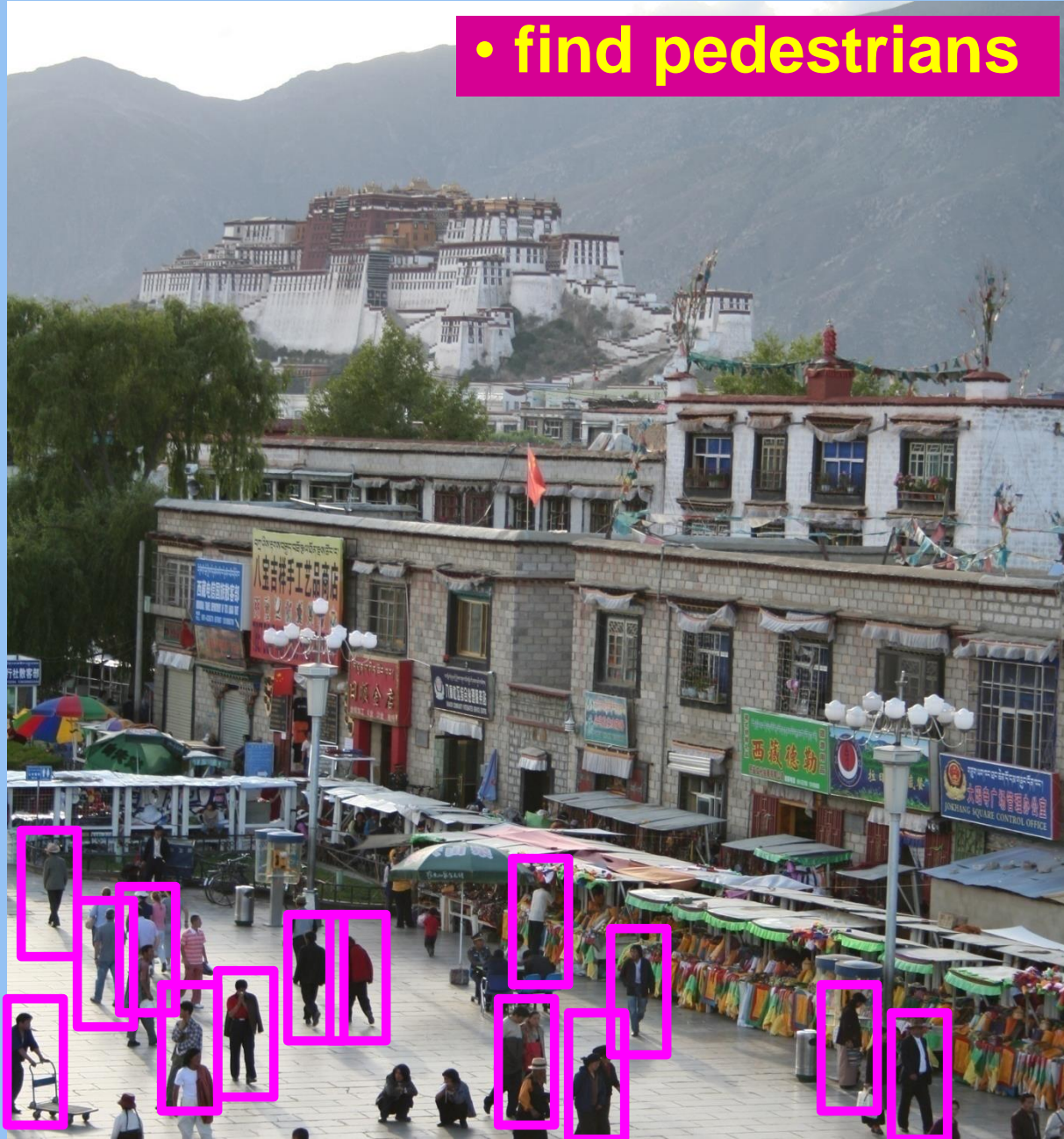
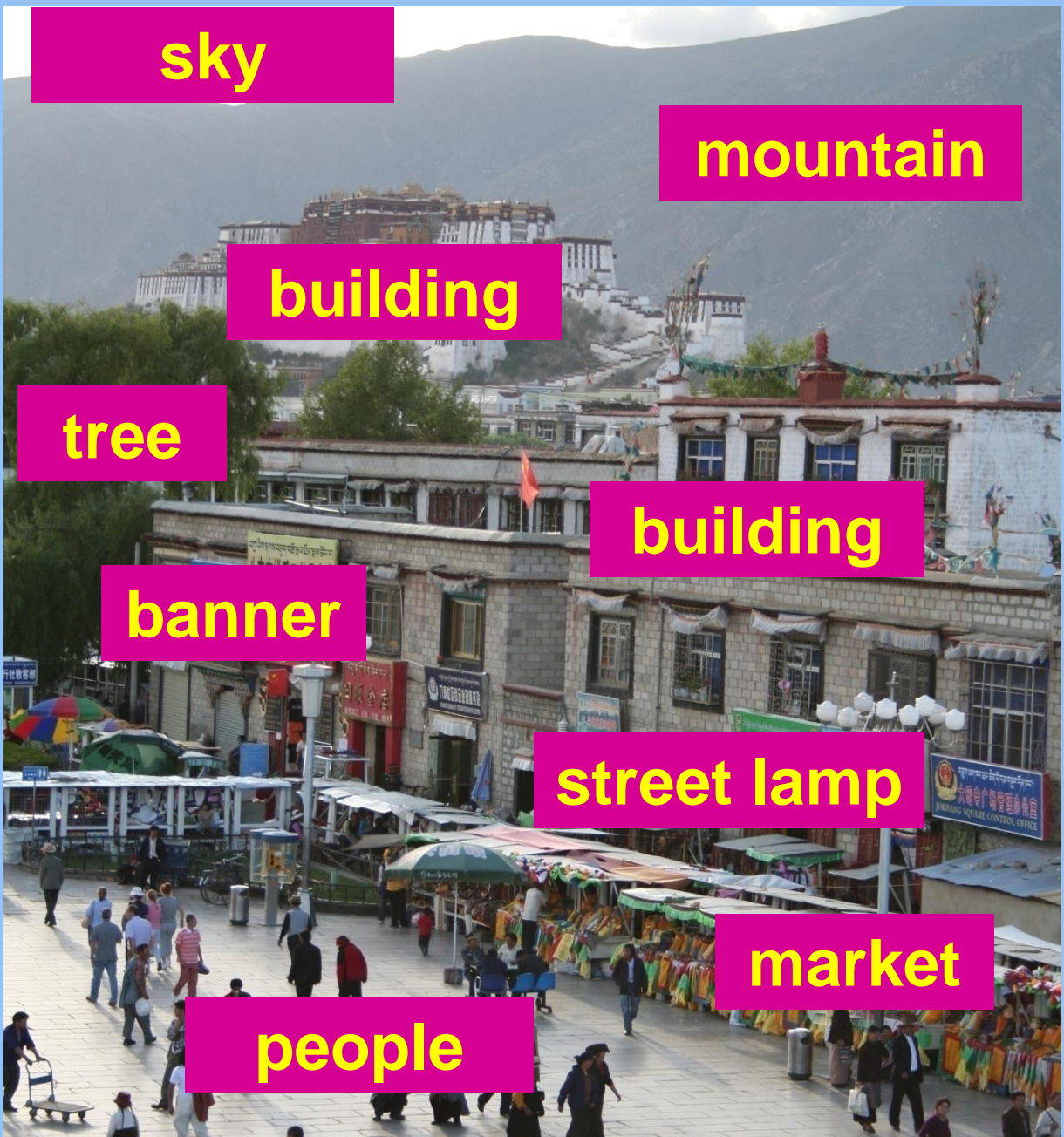


Image parsing / semantic segmentation



sky

mountain

building

tree

building

banner

street lamp

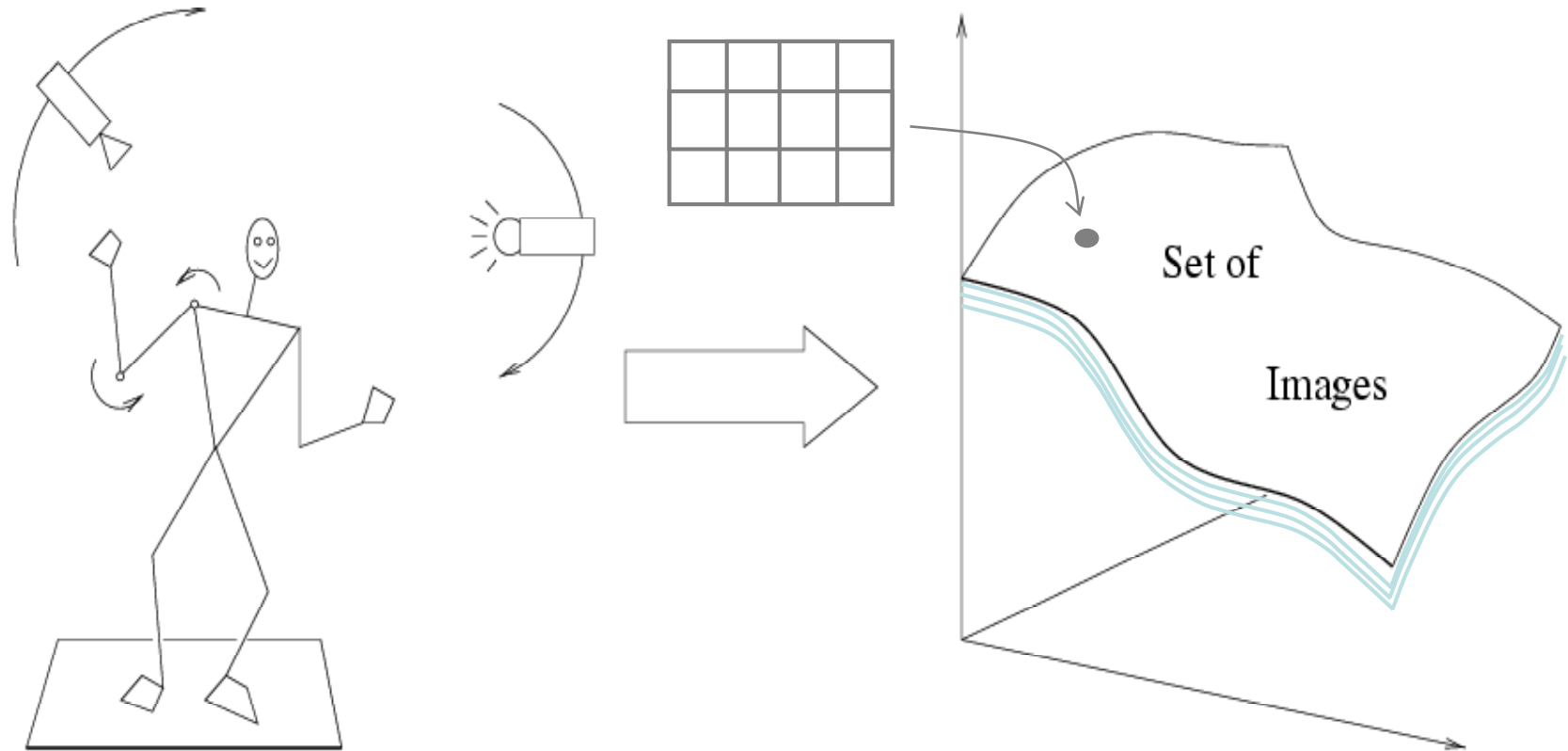
market

people

Scene understanding?



Recognition is all about modeling variability



Variability: Camera position
Illumination
Shape parameters



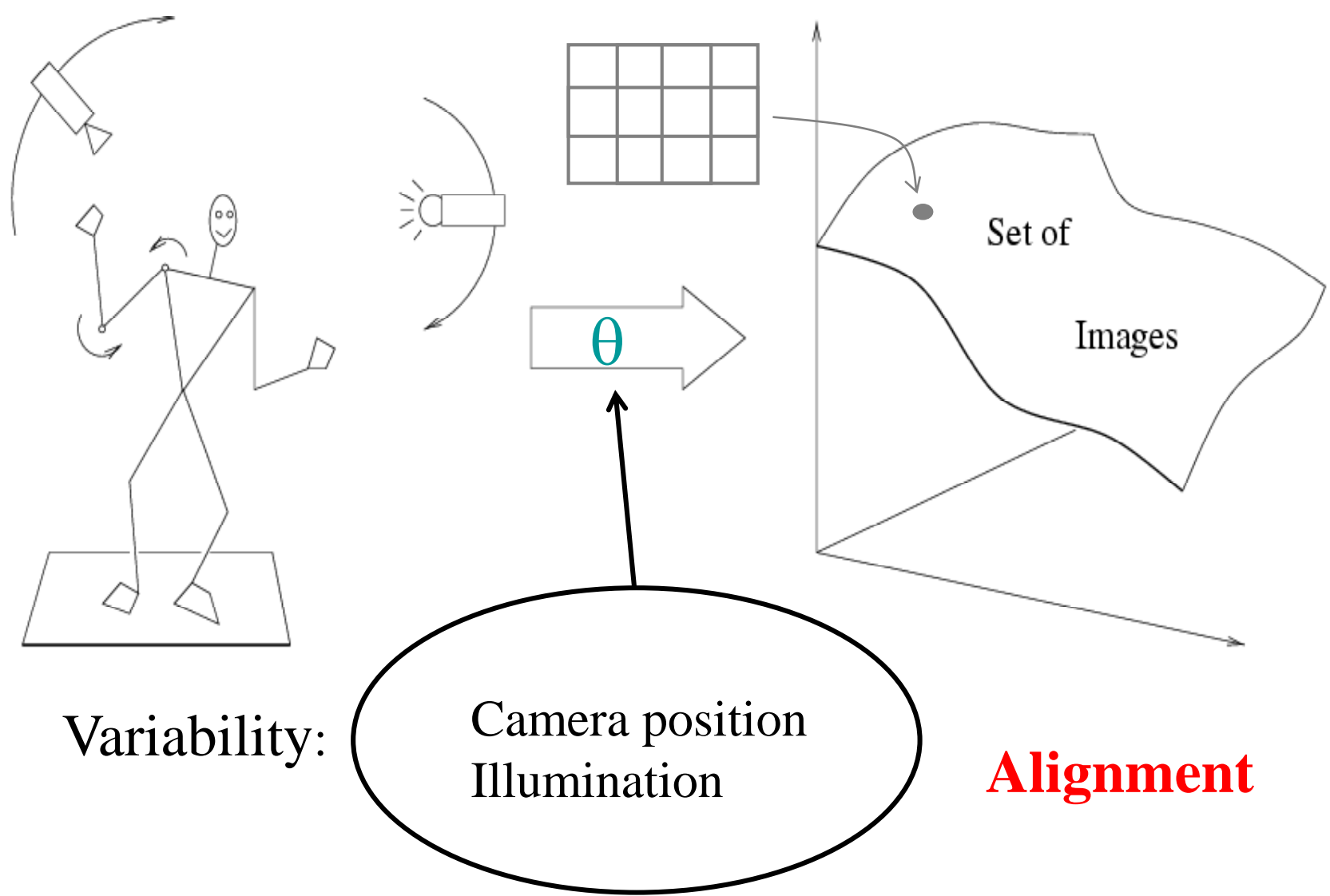
Within-class variations?

Within-class variations



History of ideas in recognition

- 1960s – early 1990s: the geometric era

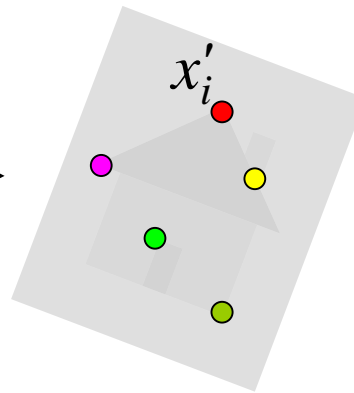
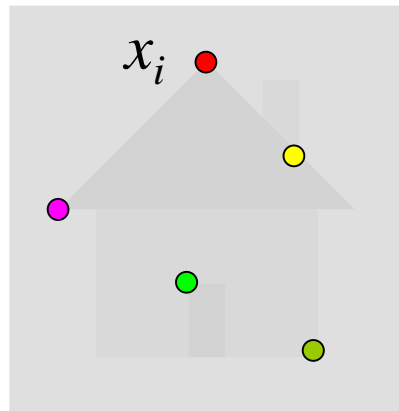


Shape: assumed known

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

Recall: Alignment

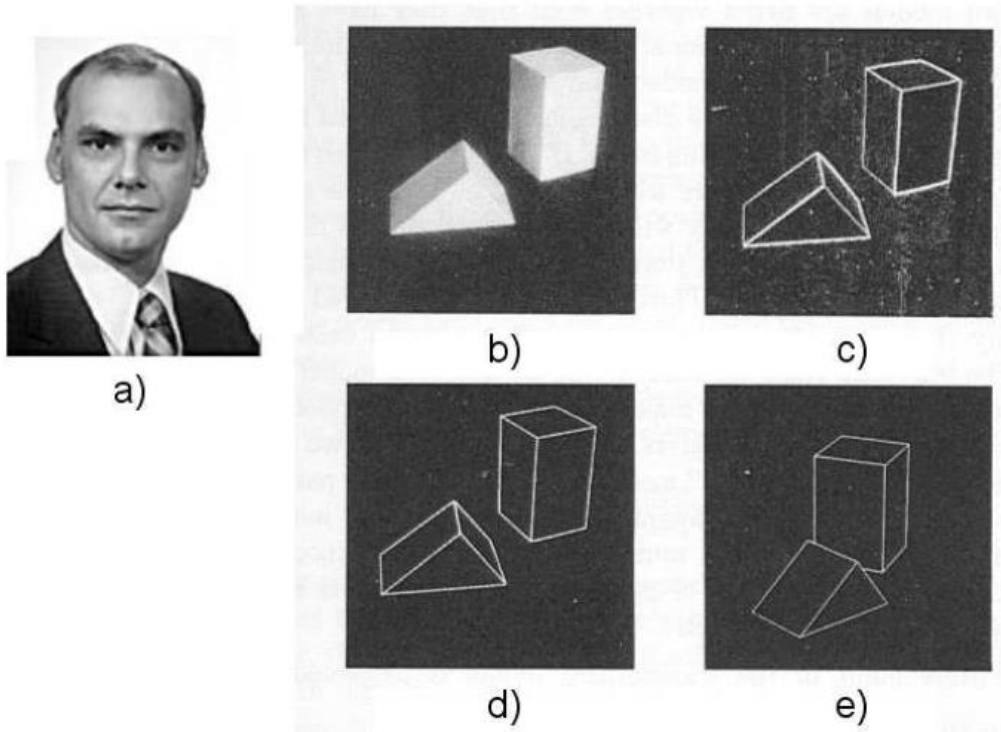
- Alignment: fitting a model to a transformation between pairs of features (*matches*) in two images



Find transformation T
that minimizes

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

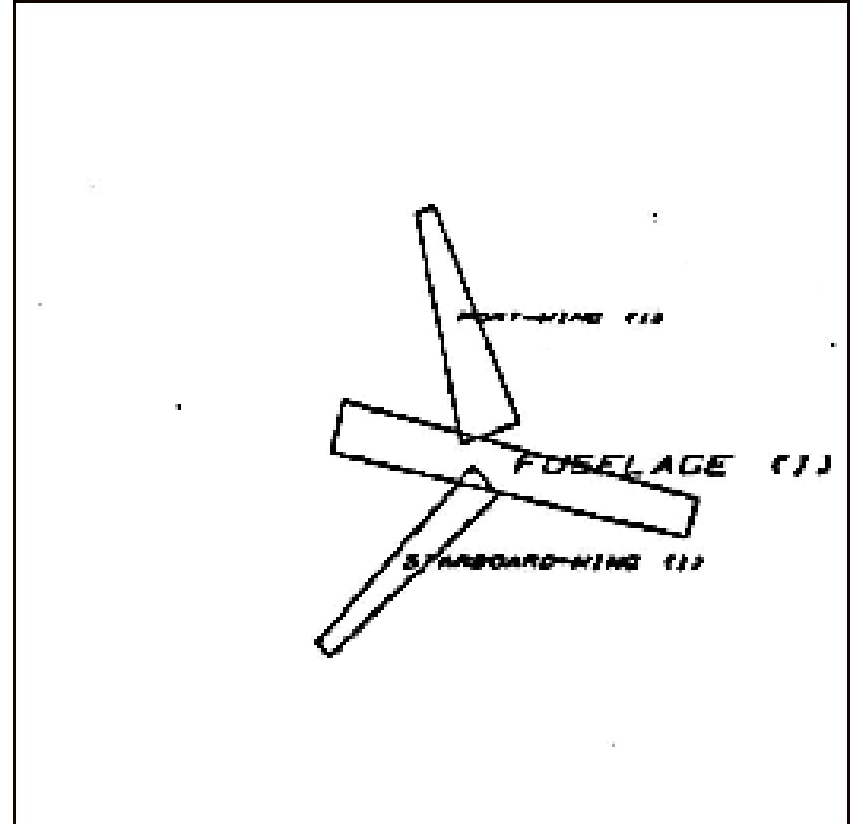
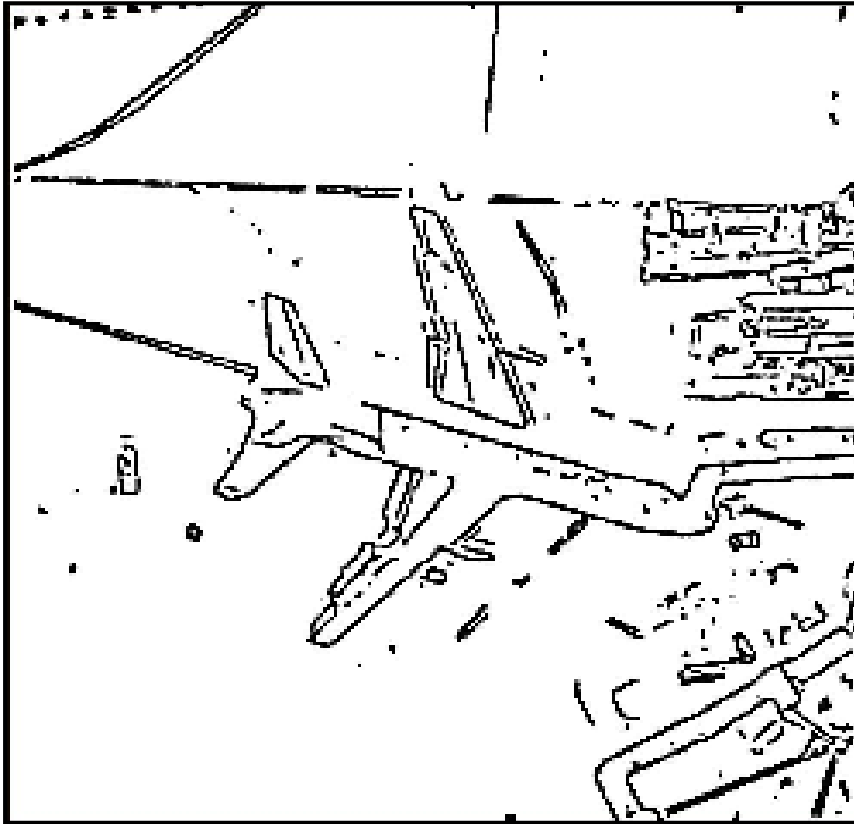
Recognition as an alignment problem: Block world



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.)

Representing and recognizing object categories is harder...



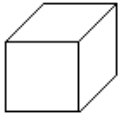
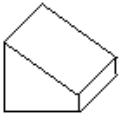
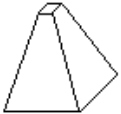

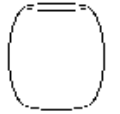
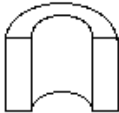

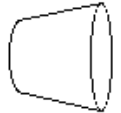


ACRONYM (Brooks and Binford, 1981)

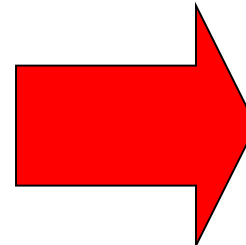
Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

Recognition by components

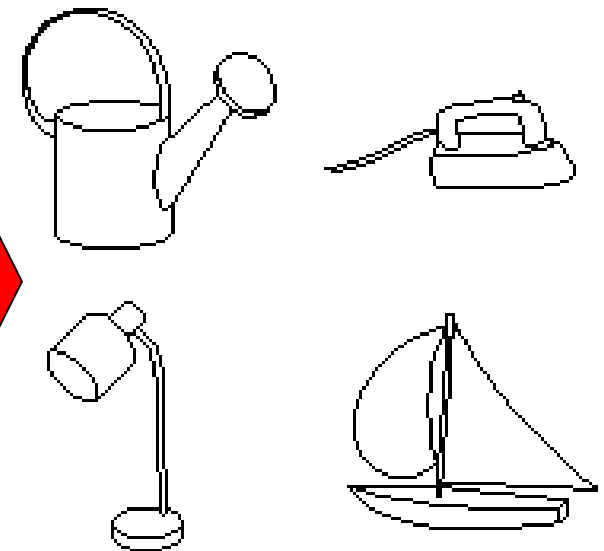
Biederman (1987)

Primitives (geons)

<p>Cube</p>  <p>Straight Edge Straight Axis Constant</p>	<p>Wedge</p>  <p>Straight Edge Straight Axis Expanded</p>	<p>Pyramid</p>  <p>Straight Edge Straight Axis Expanded</p>	<p>Cylinder</p>  <p>Curved Edge Straight Axis Constant</p>	<p>Barrel</p>  <p>Curved Edge Straight Axis Exp & Cont</p>
<p>Arch</p>  <p>Straight Edge Curved Axis Constant</p>	<p>Cone</p>  <p>Curved Edge Straight Axis Expanded</p>	<p>Expanded Cylinder</p>  <p>Curved Edge Straight Axis Expanded</p>	<p>Handle</p>  <p>Curved Edge Curved Axis Constant</p>	<p>Expanded Handle</p>  <p>Curved Edge Curved Axis Expanded</p>

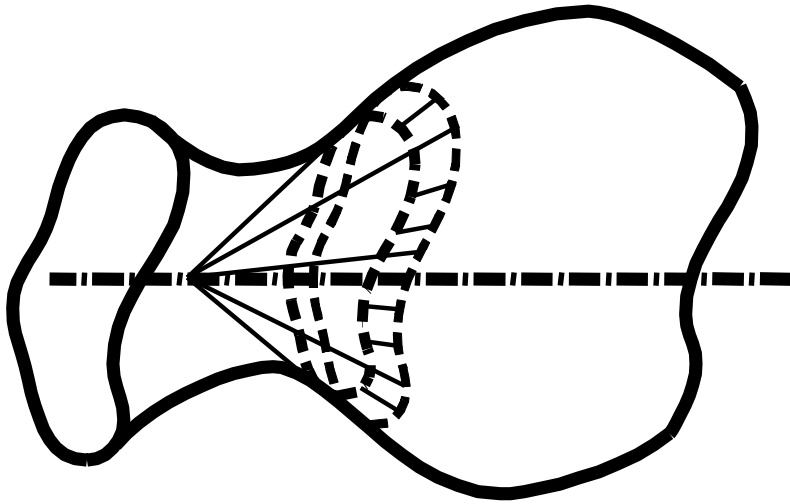


Objects

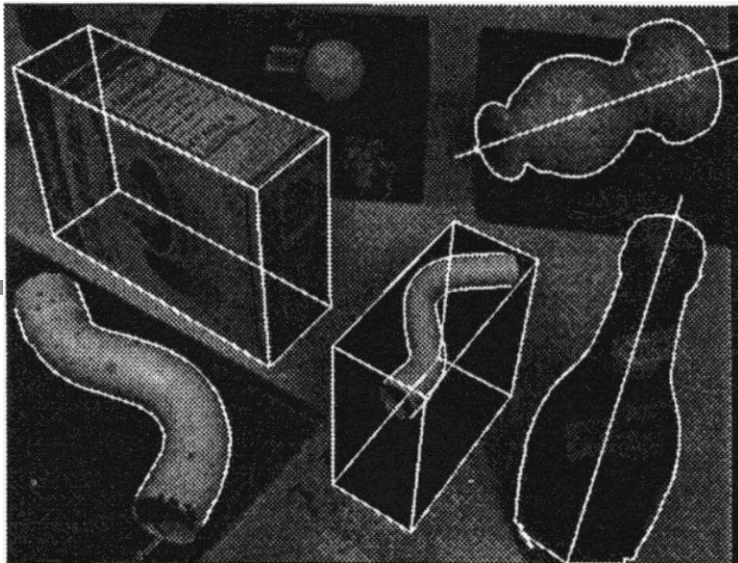


http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

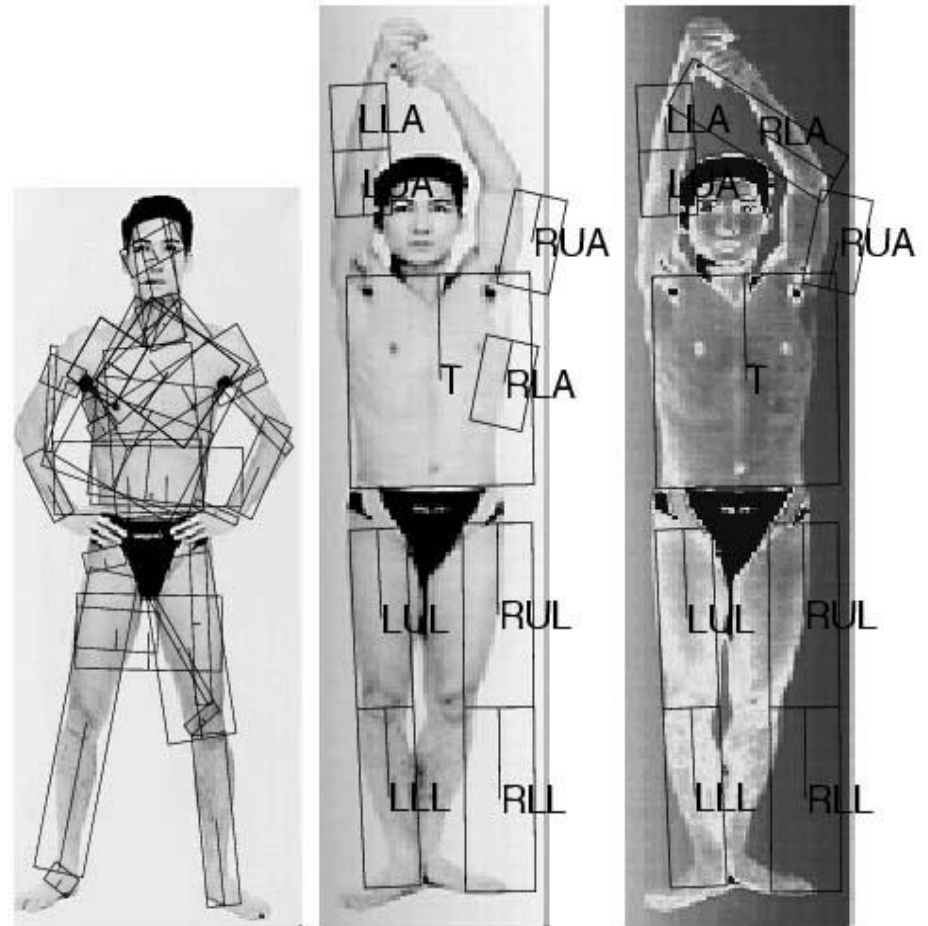
General shape primitives?



Generalized cylinders
Ponce et al. (1989)



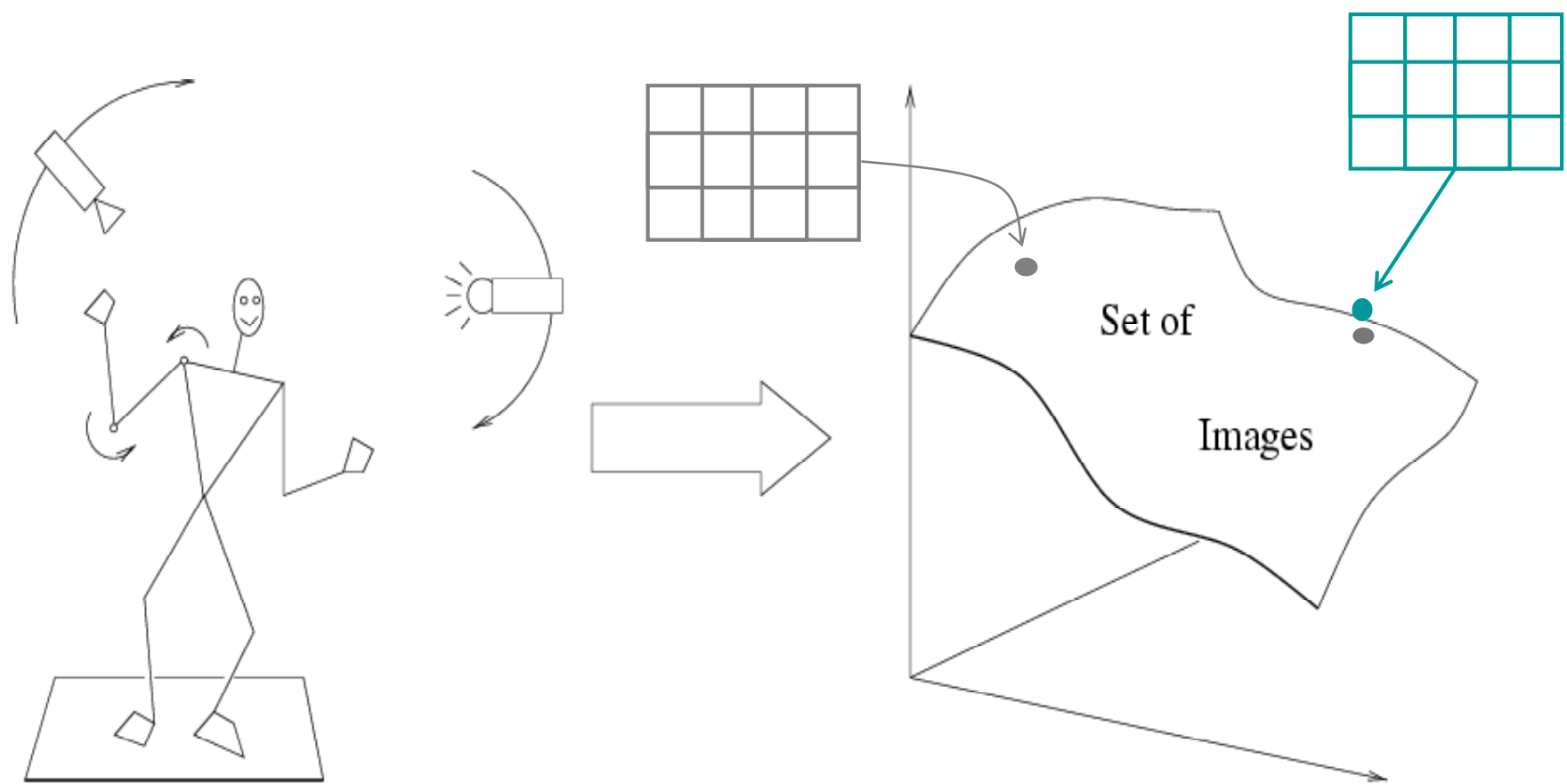
Zisserman et al. (1995)



Forsyth (2000)

History of ideas in recognition

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

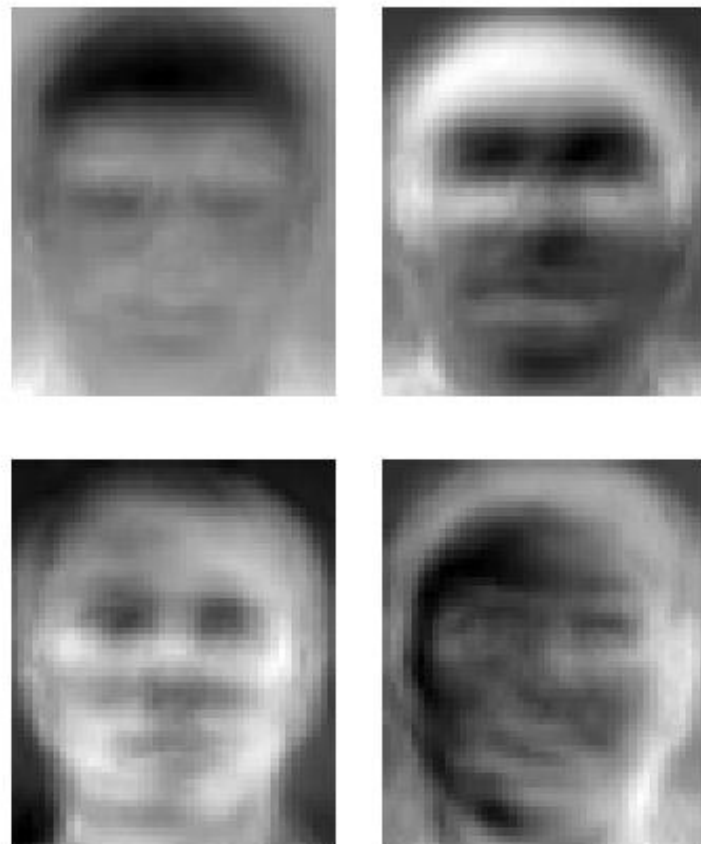


Empirical models of image variability

Appearance-based techniques

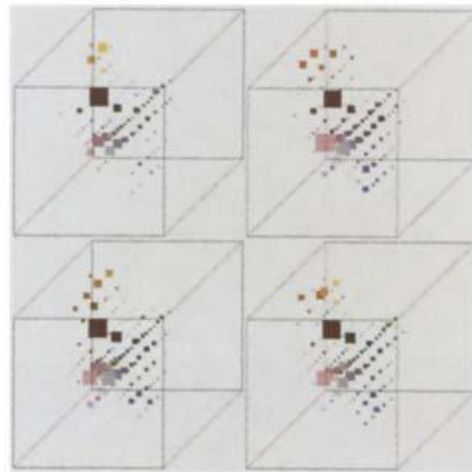
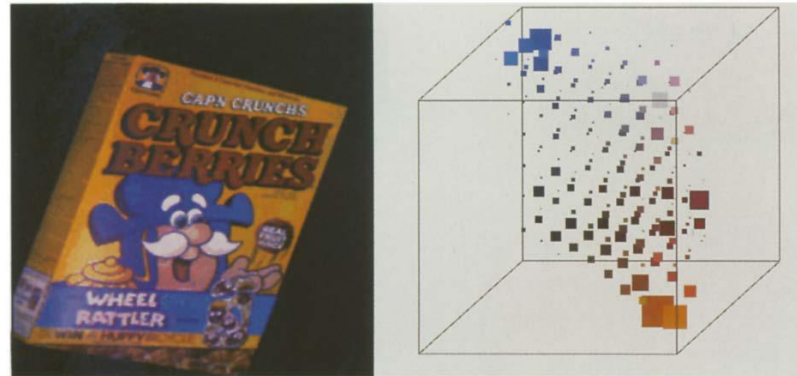
Turk & Pentland (1991); Murase & Nayar (1995); etc.

Eigenfaces (Turk & Pentland, 1991)



Experimental Condition	Correct/Unknown Recognition Percentage		
	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

Color Histograms



Swain and Ballard, [Color Indexing](#), IJCV 1991.

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



- 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

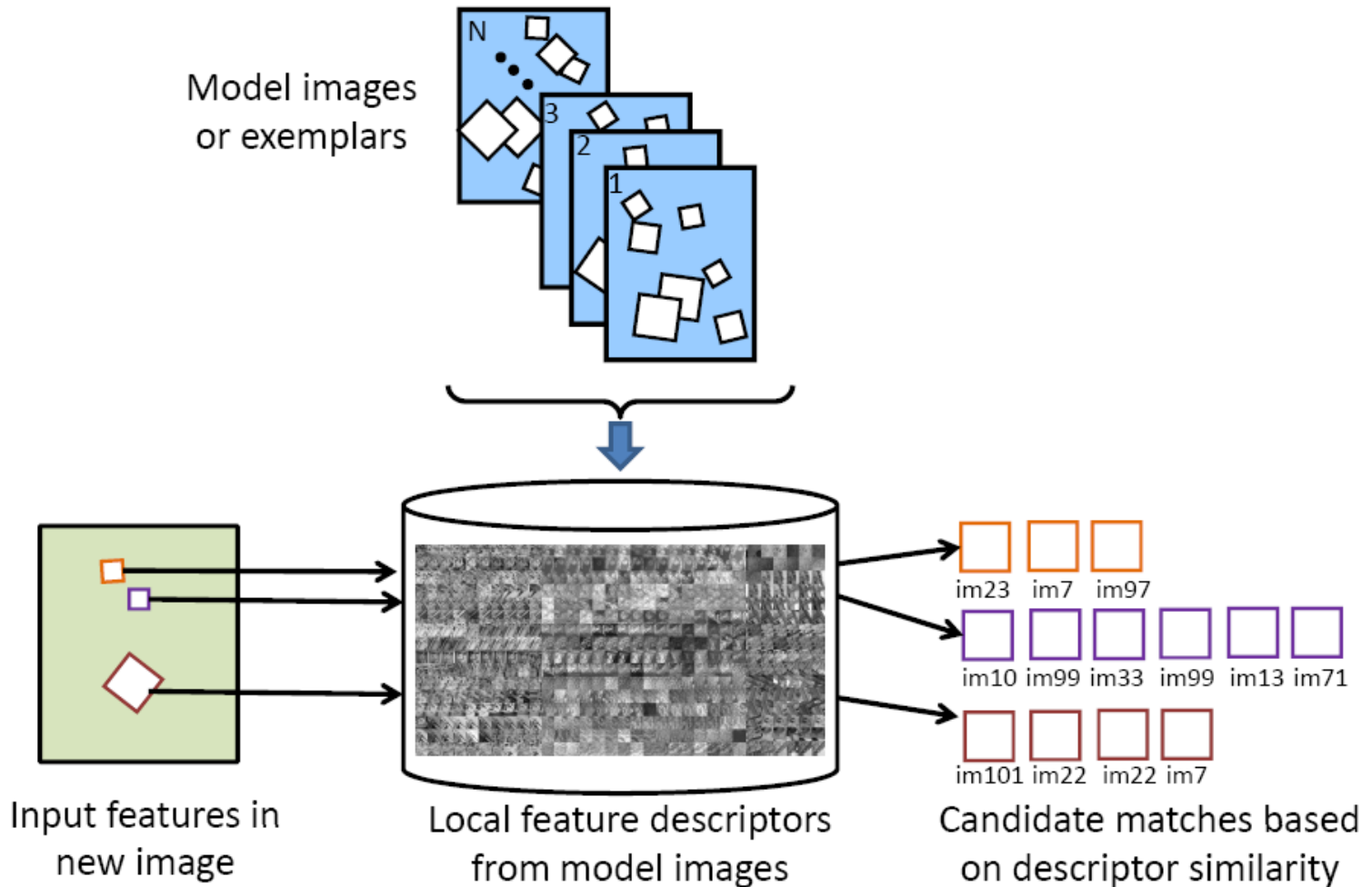
Local features for object instance recognition



D. Lowe (1999, 2004)

Large-scale image search

Combining local features, indexing, and spatial constraints



Large-scale image search

Combining local features, indexing, and spatial constraints



Large-scale image search

Combining local features, indexing, and spatial constraints

Google Goggles in Action

Click the icons below to see the different ways Google Goggles can be used.



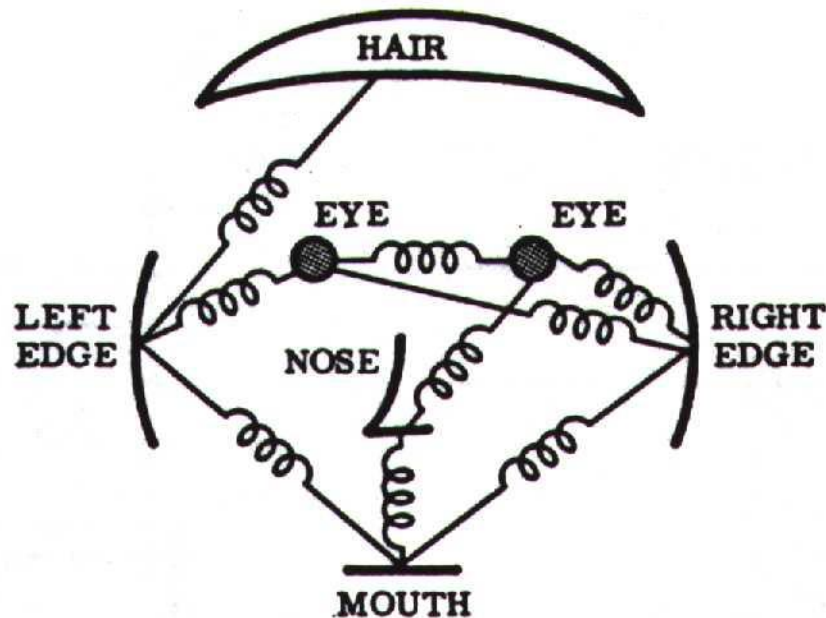
Available on phones that run Android 1.6+ (i.e. Donut or Eclair)

History of ideas in recognition

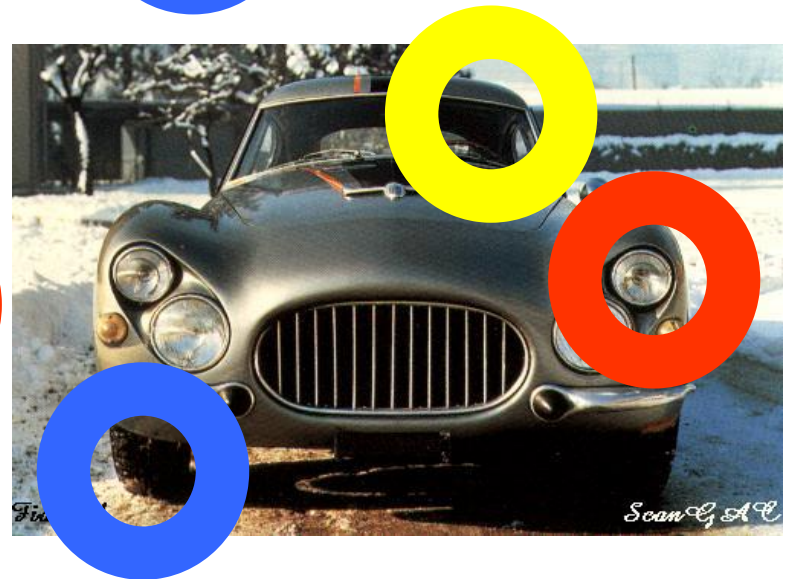
- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models

Parts-and-shape models

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



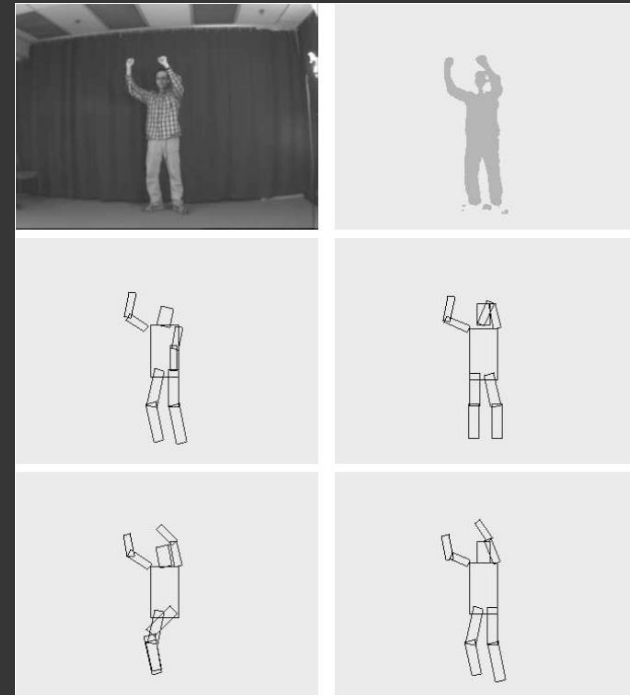
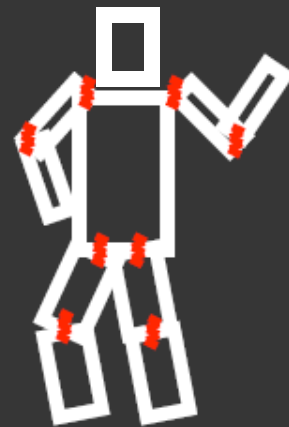
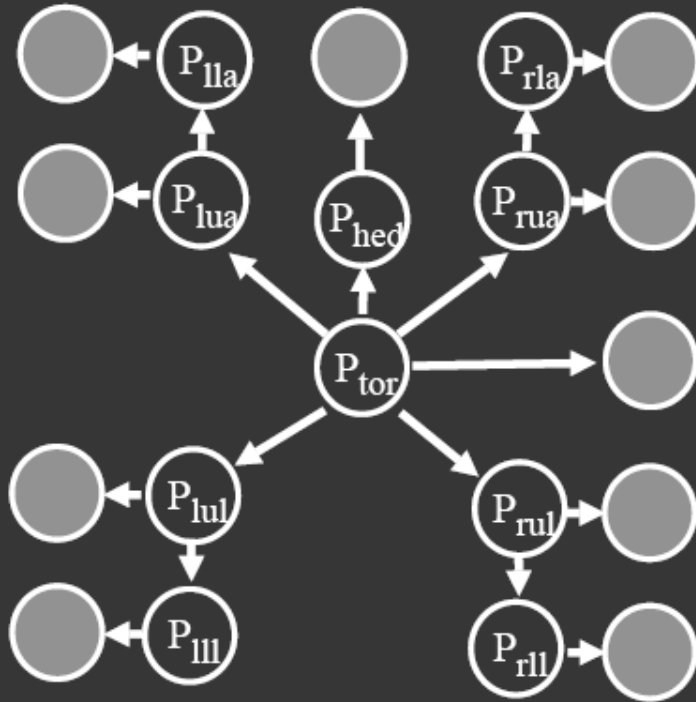
Constellation models



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

Pictorial structure model

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

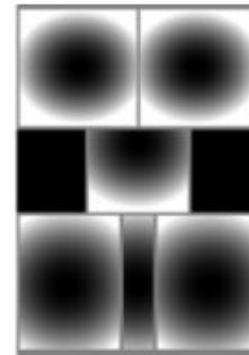
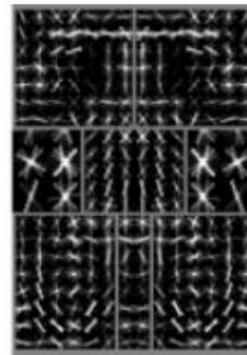
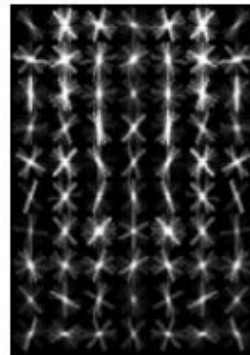
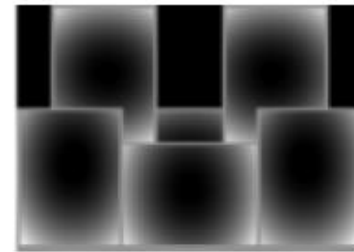
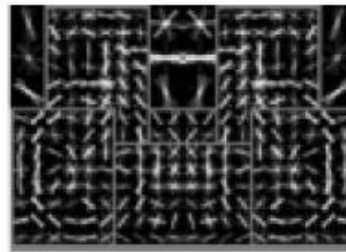
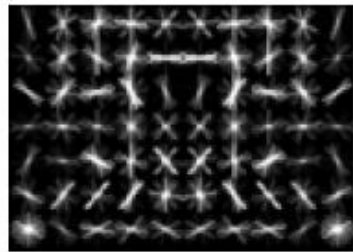
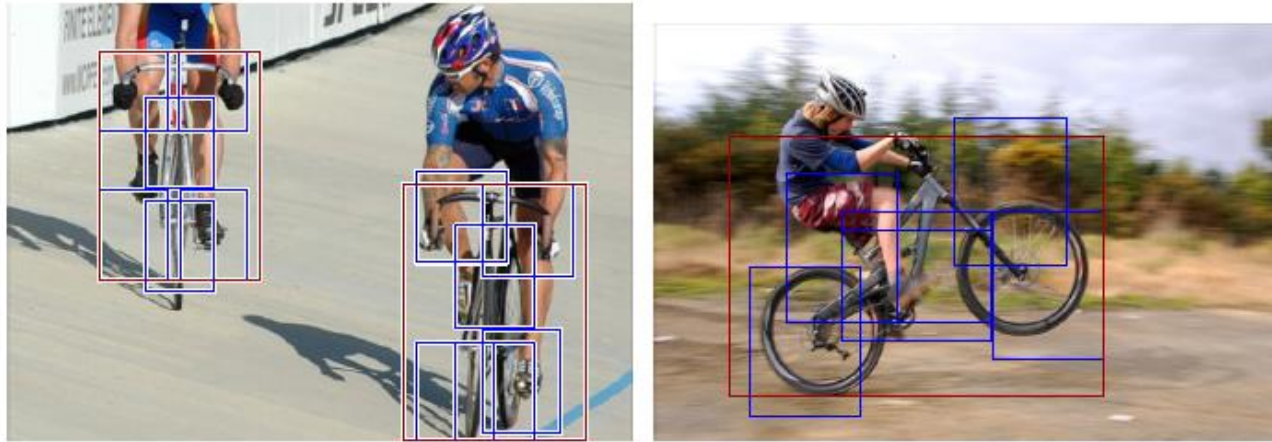


$$\Pr(P_{\text{tor}}, P_{\text{arm}}, \dots | \text{Im}) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(\text{Im}(P_i))$$

↑
↑

part geometry
part appearance

Discriminatively trained part-based models

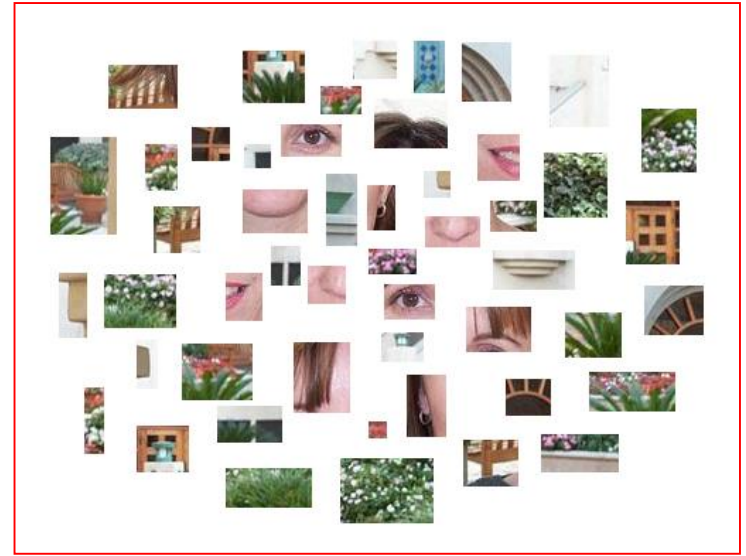
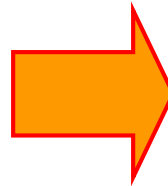


P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, ["Object Detection with Discriminatively Trained Part-Based Models,"](#) PAMI 2009

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

Bag-of-features models



Bag-of-features models

Object

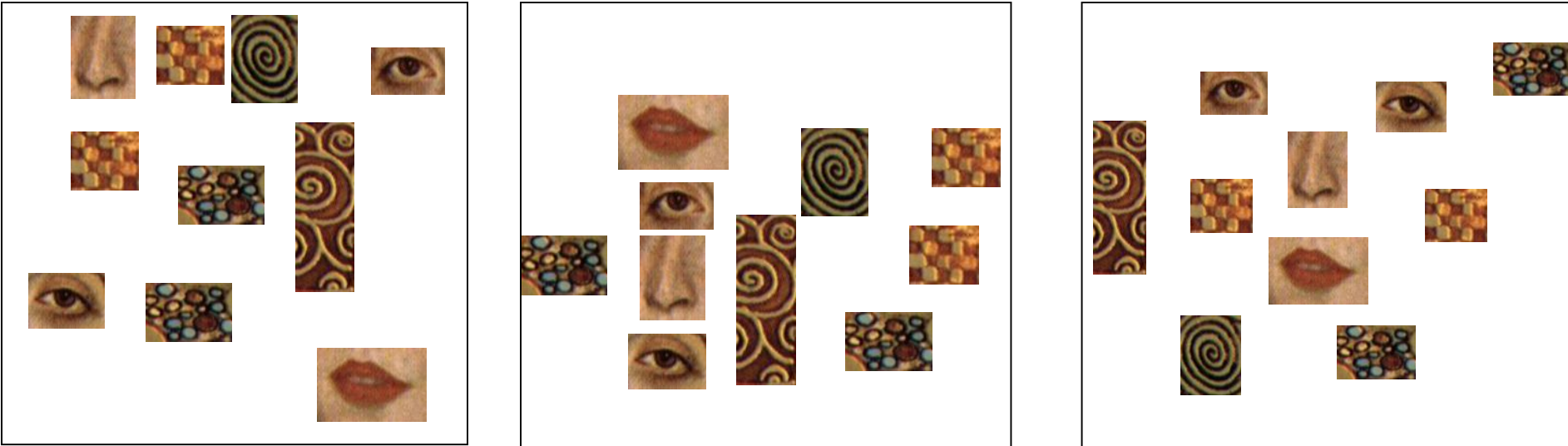


**Bag of
'words'**



Objects as texture

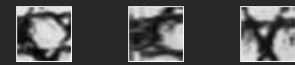
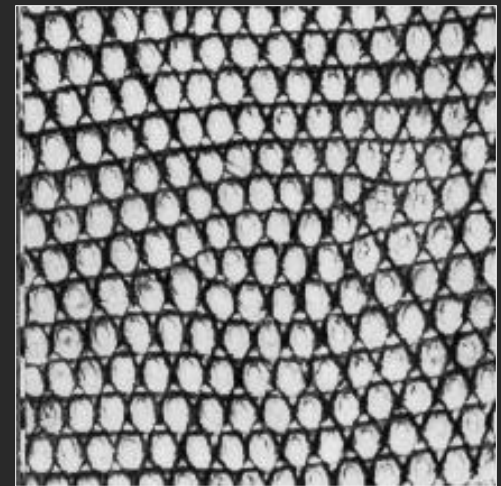
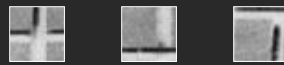
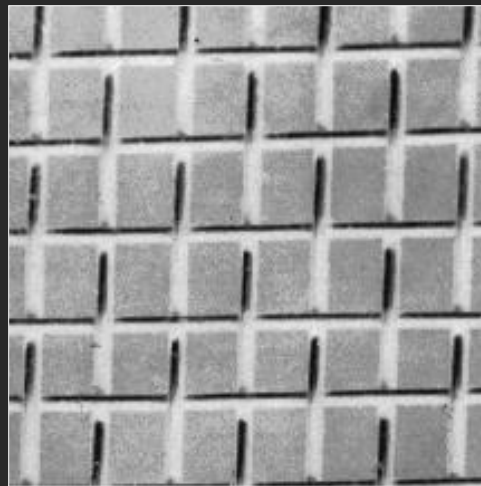
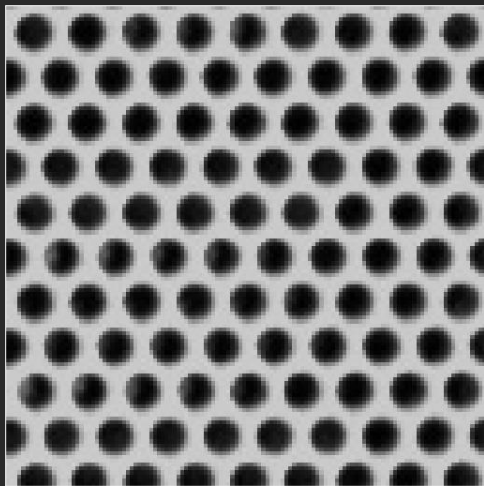
- All of these are treated as being the same



- No distinction between foreground and background: scene recognition?

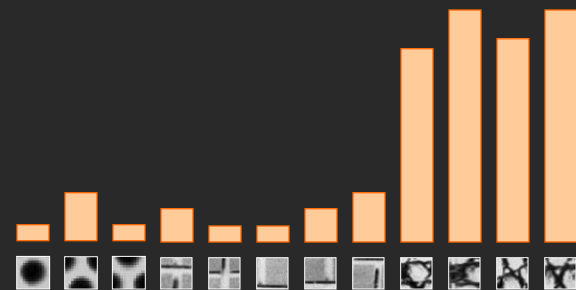
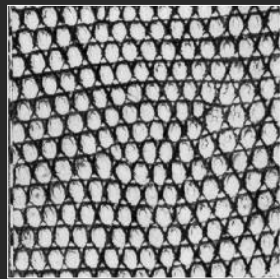
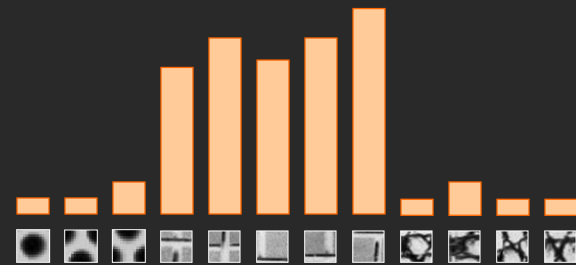
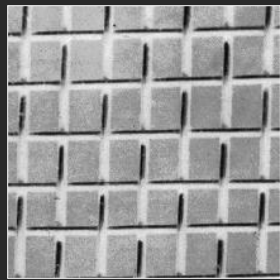
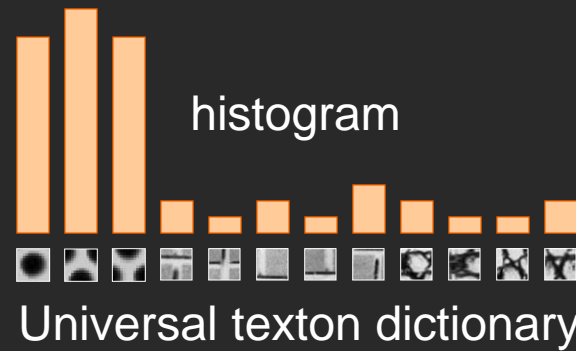
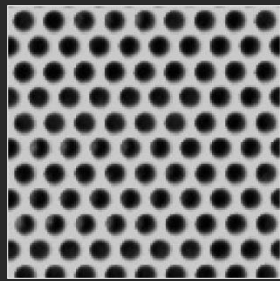
Origin 1: Texture recognition

- 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



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



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

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)

2007-01-23: State of the Union Address

George W. Bush (2001-)

abandon accountable affordable afghanistan africa aided ally anbar armed army **baghdad** bless **challenges** chamber chaos
choices civilians coalition commanders **commitment** confident confront congressman constitution corps debates deduction
deficit deliver **democratic** deploy dikembe diplomacy disruptions earmarks **economy** einstein **elections** eliminates
expand **extremists** failing faithful families **freedom** fuel **funding** god haven ideology immigration impose
insurgents iran **iraq** islam julie lebanon love madam marine math medicare moderation neighborhoods nuclear offensive
palestinian payroll province pursuing **qaeda** radical regimes resolve retreat rieman sacrifices science sectarian senate
september **shia** stays strength students succeed sunni **tax** territories **terrorists** threats uphold victory
violence violent **war** washington weapons wesley

Origin 2: Bag-of-words models

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



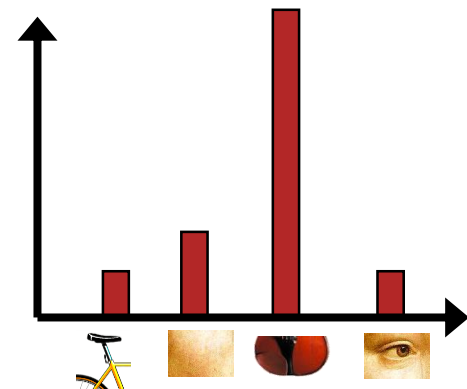
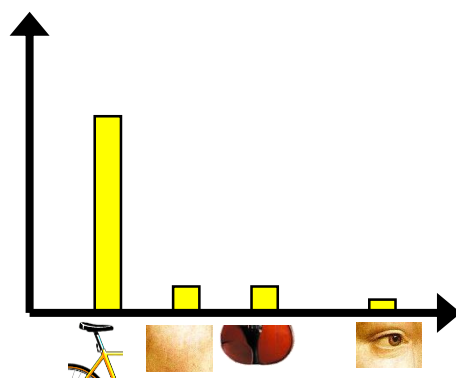
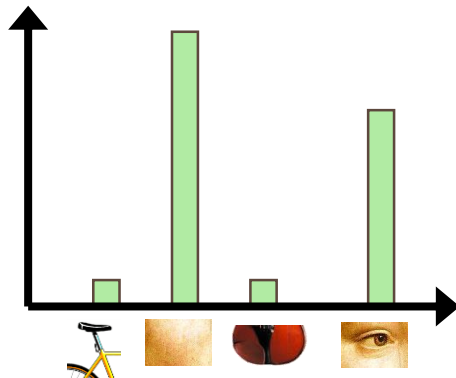
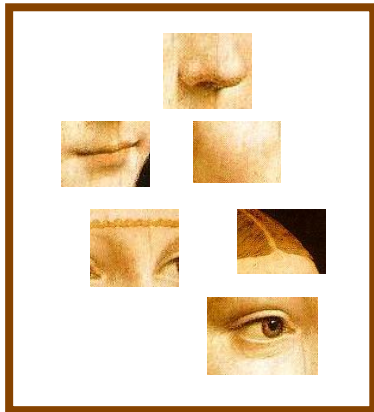
Origin 2: Bag-of-words models

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



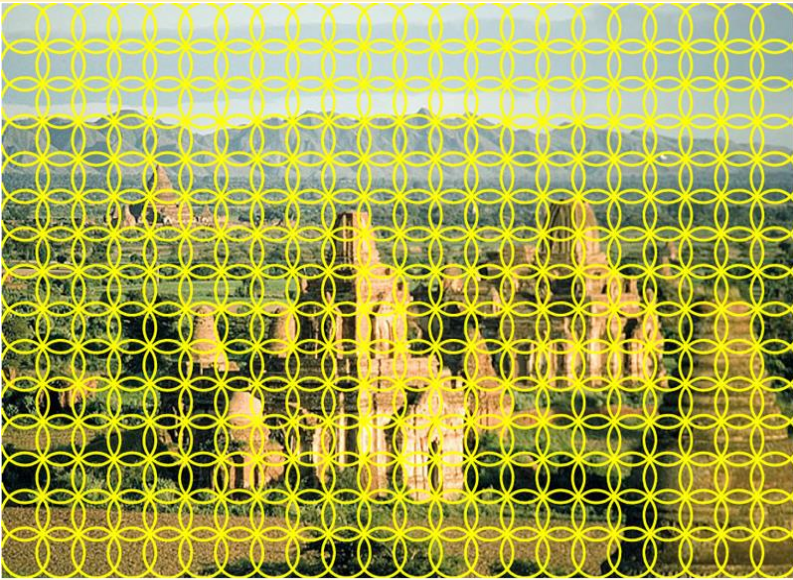
Bag-of-features steps

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

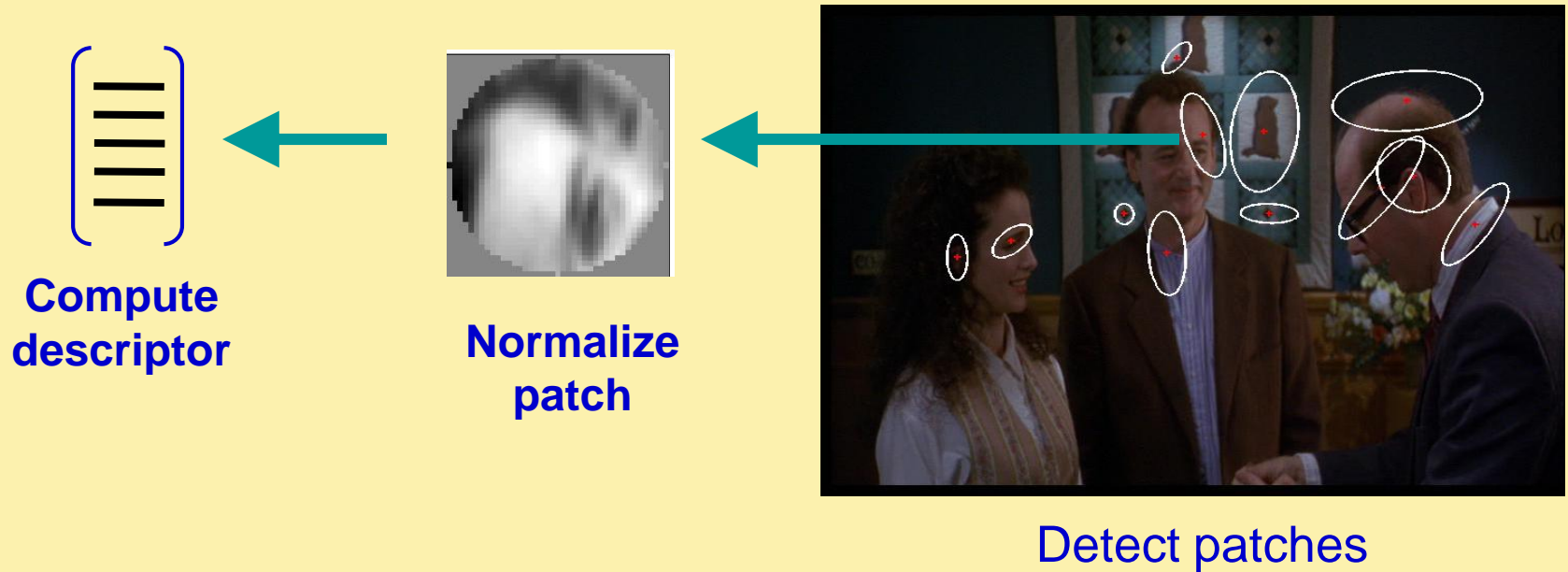


1. Feature extraction

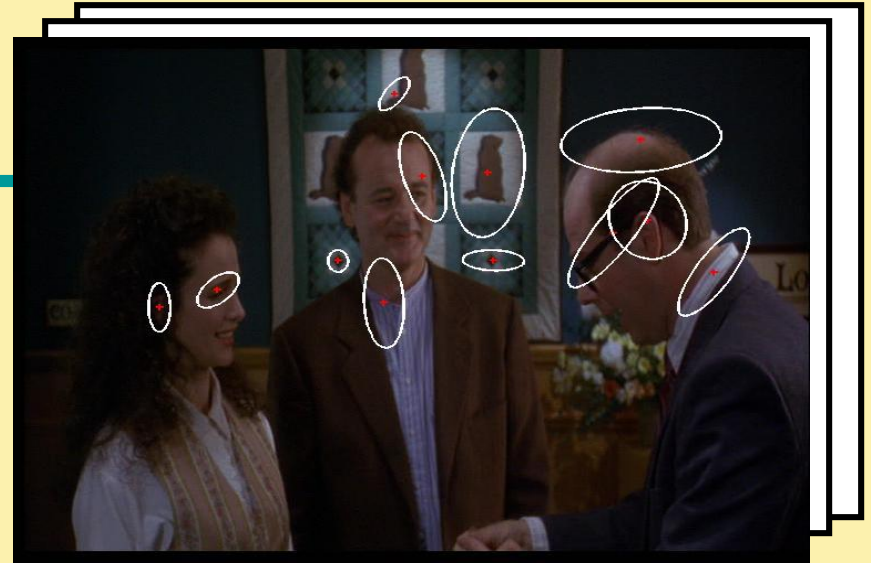
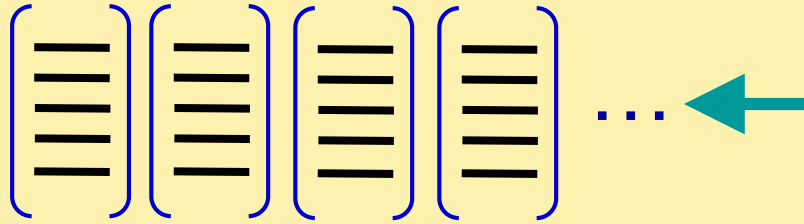
- Regular grid or interest regions



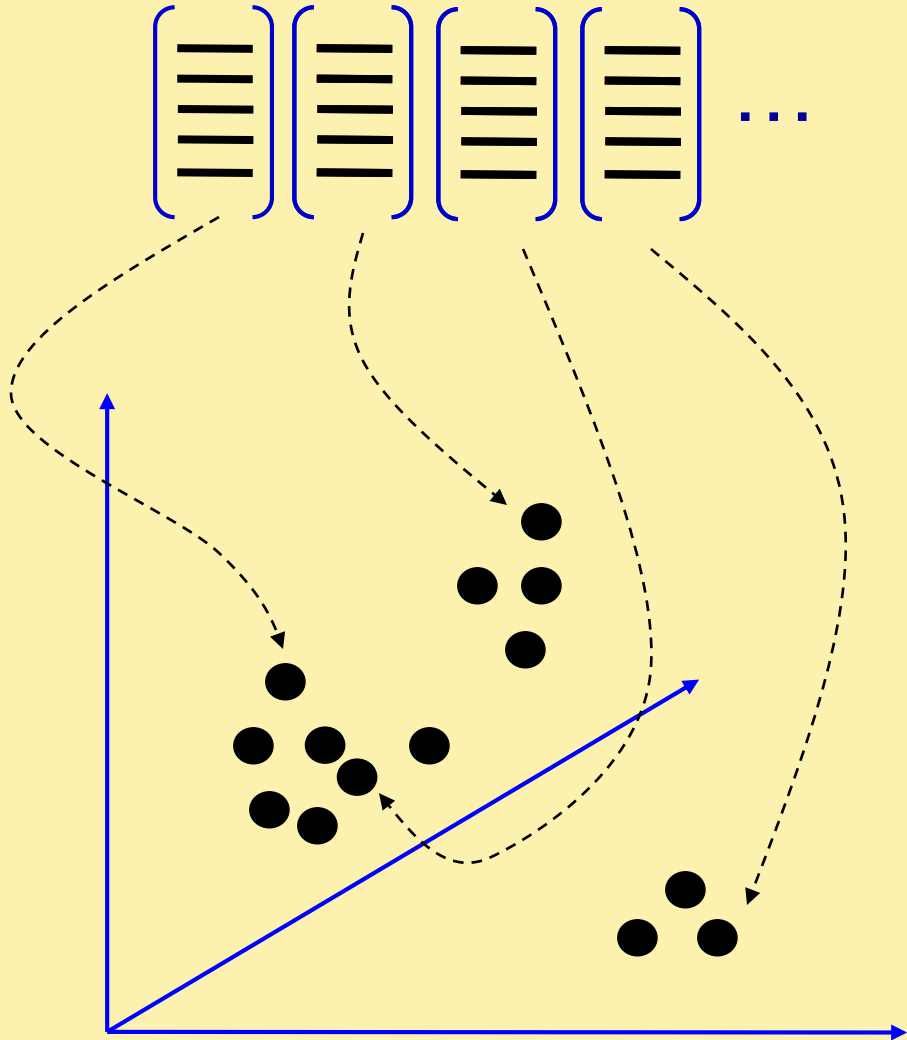
1. Feature extraction



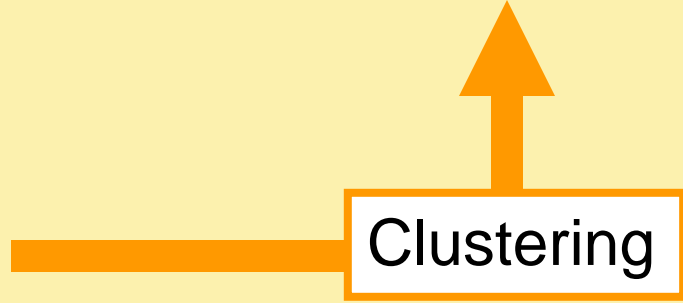
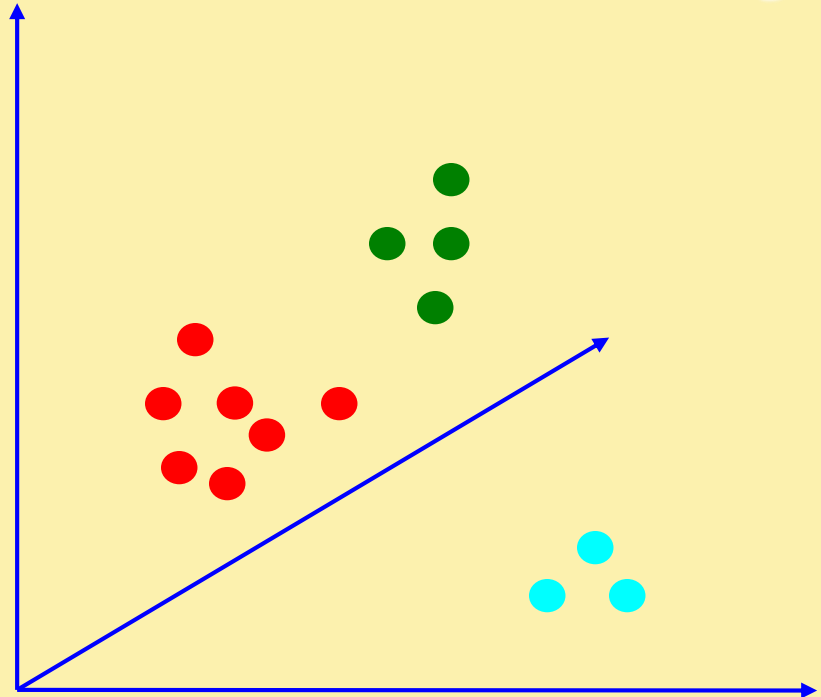
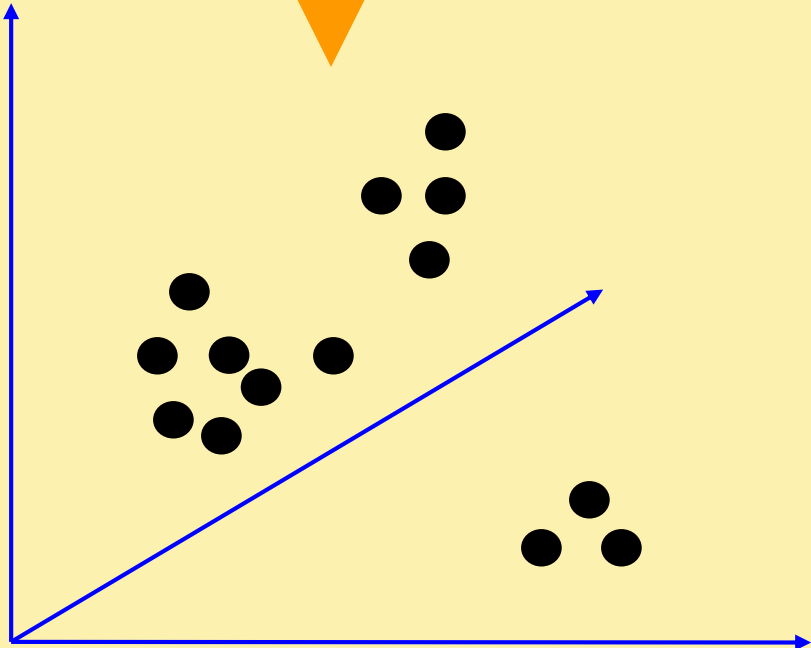
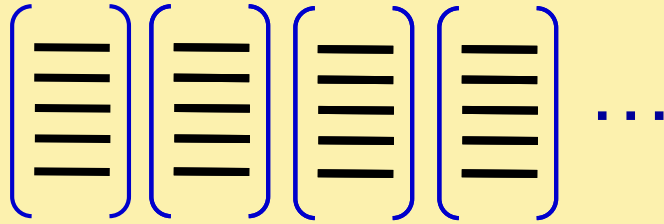
1. Feature extraction



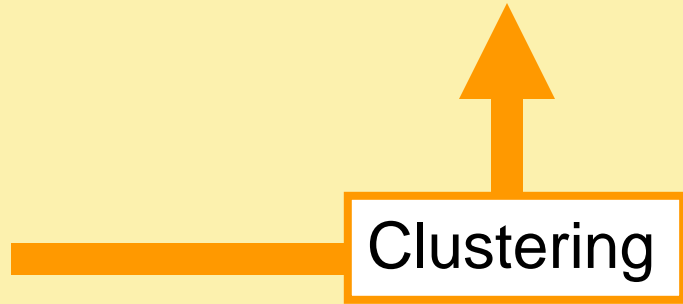
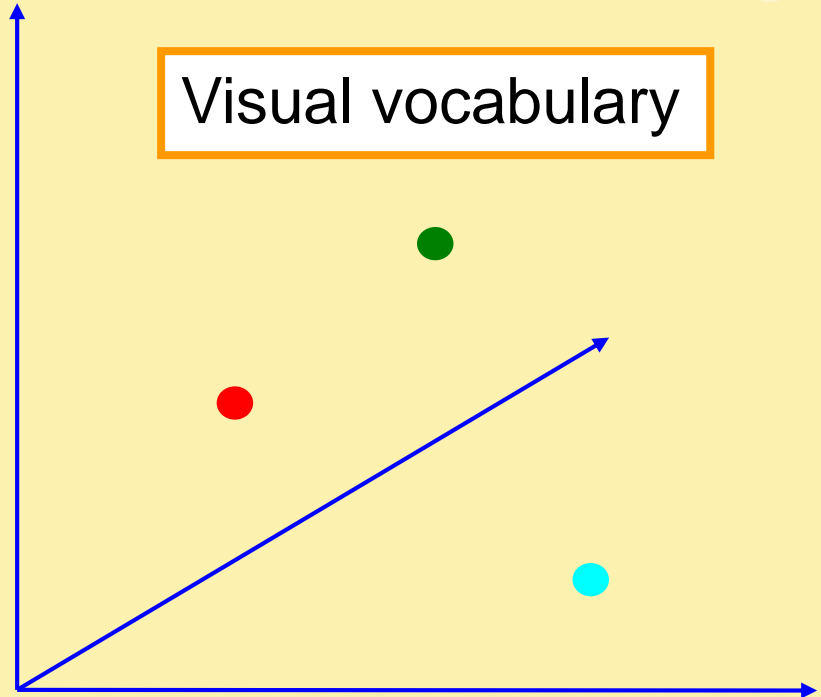
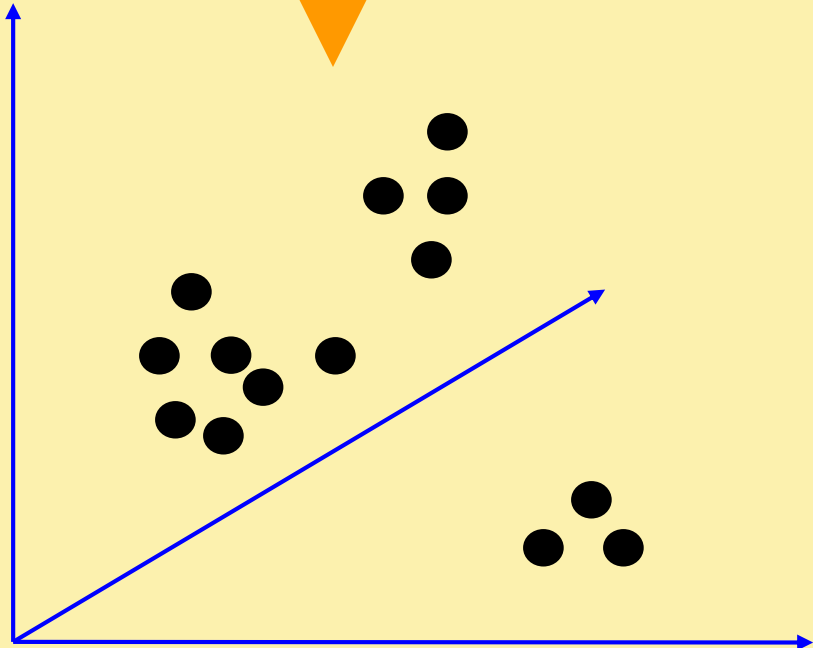
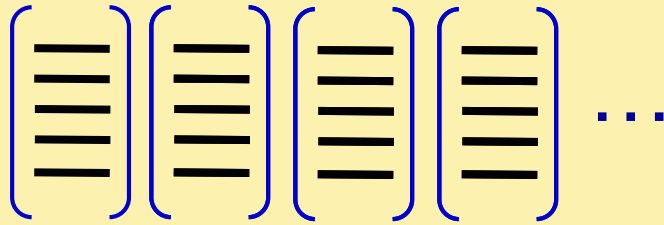
2. Learning the visual vocabulary



2. Learning the visual vocabulary



2. Learning the visual vocabulary



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:

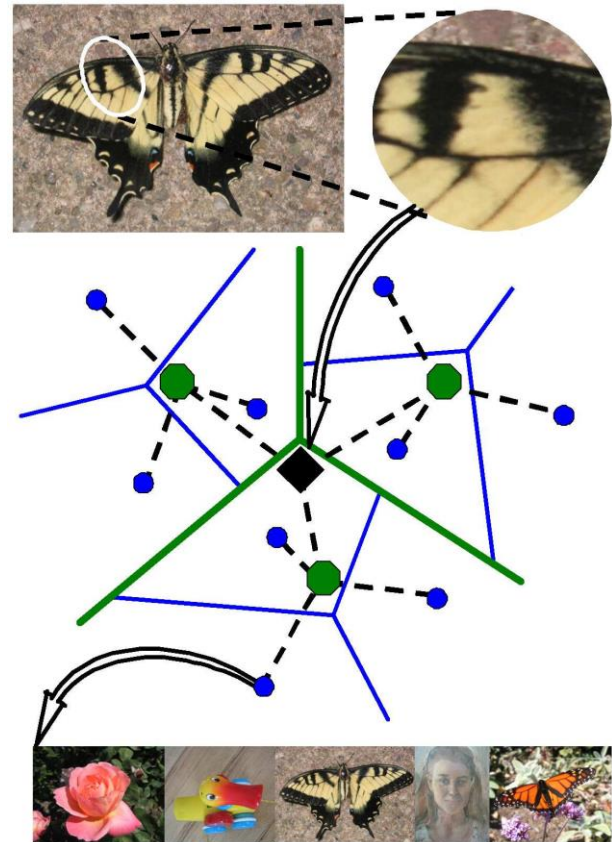
- Randomly initialize K cluster centers
- Iterate until convergence:
 - Assign each data point to the nearest center
 - Recompute each cluster center as the mean of all points assigned to it

Clustering and vector quantization

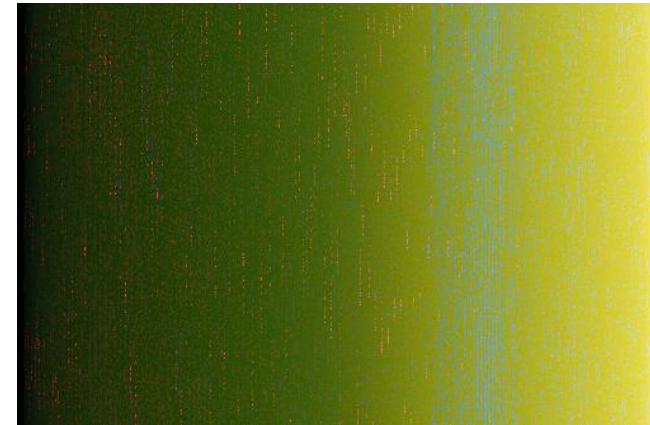
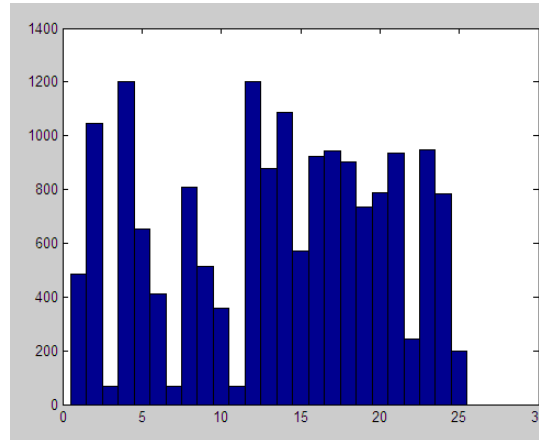
- Clustering is a common method for learning a visual vocabulary or codebook
 - Unsupervised learning process
 - Each cluster center produced by k-means becomes a codevector
 - Codebook can be learned on separate training set
 - Provided the training set is sufficiently representative, the codebook will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

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)

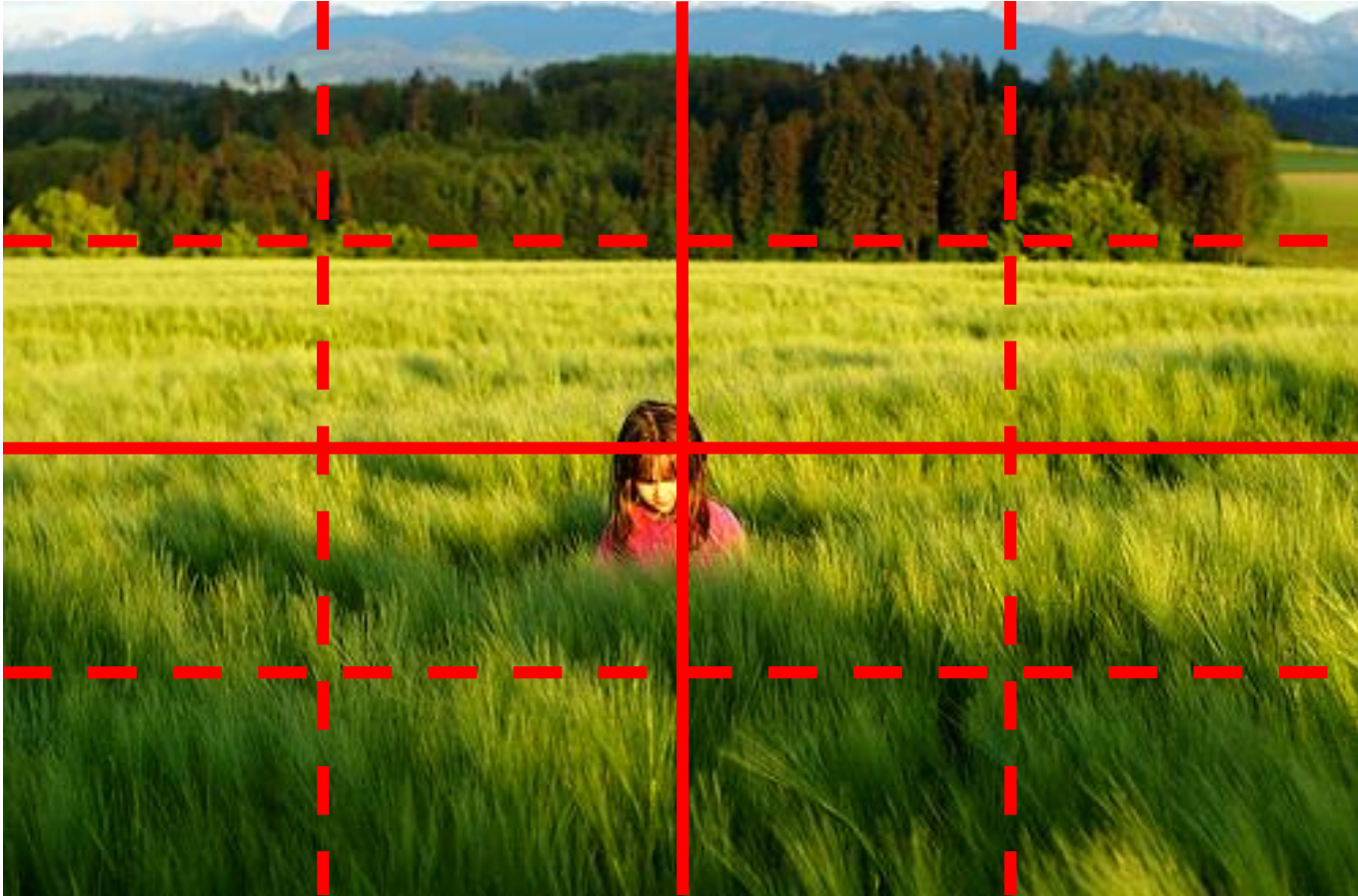


But what about layout?



All of these images have the same color histogram

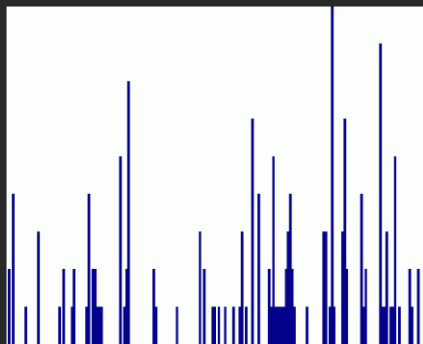
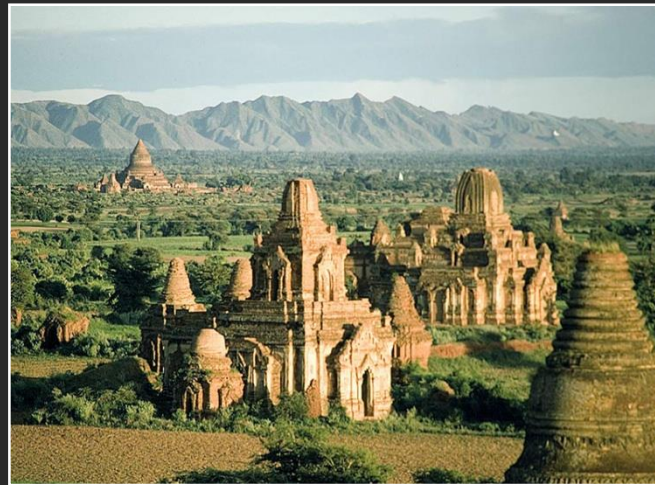
Spatial pyramid



Compute histogram in each spatial bin

Spatial pyramid representation

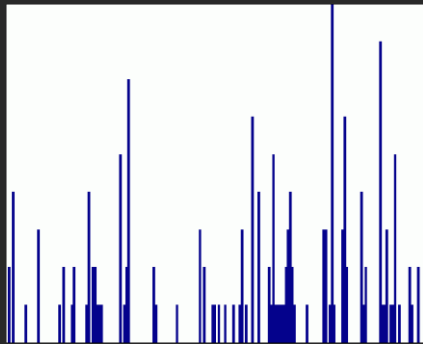
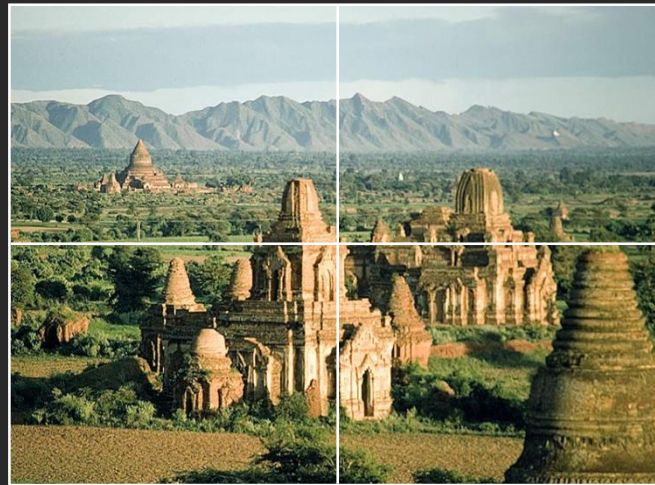
- Extension of a bag of features
- Locally orderless representation at several levels of resolution



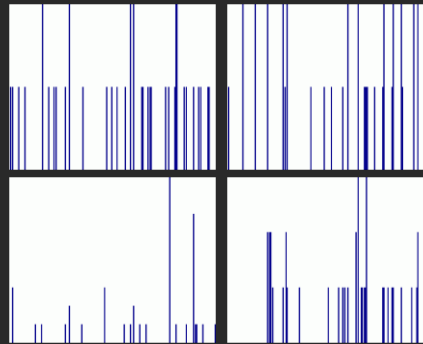
level 0

Spatial pyramid representation

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



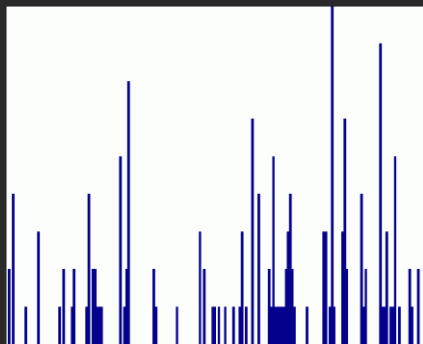
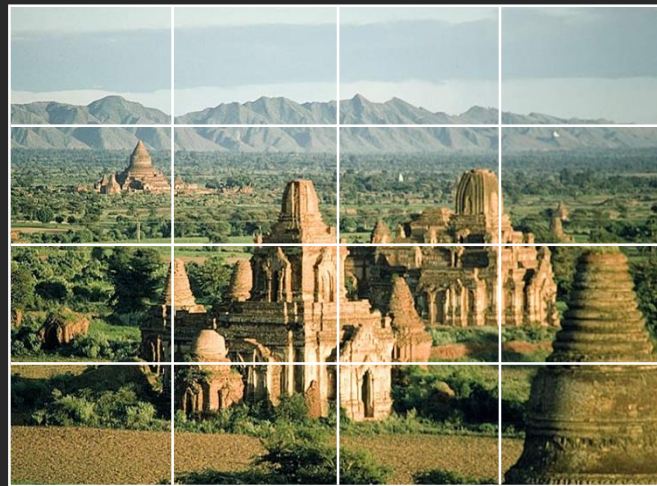
level 0



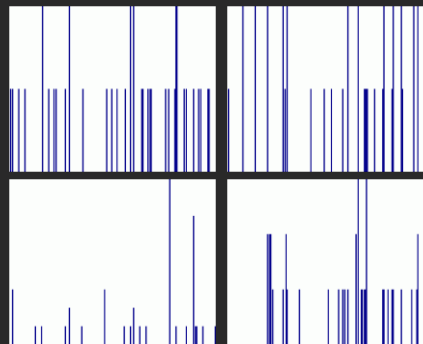
level 1

Spatial pyramid representation

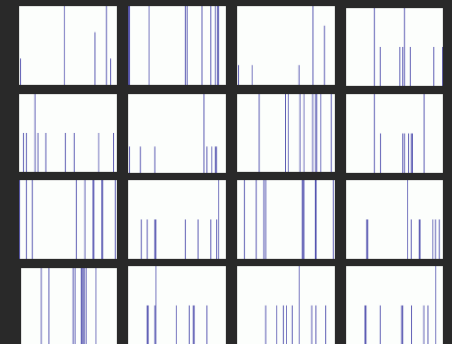
- Extension of a bag of features
- Locally orderless representation at several levels of resolution



level 0

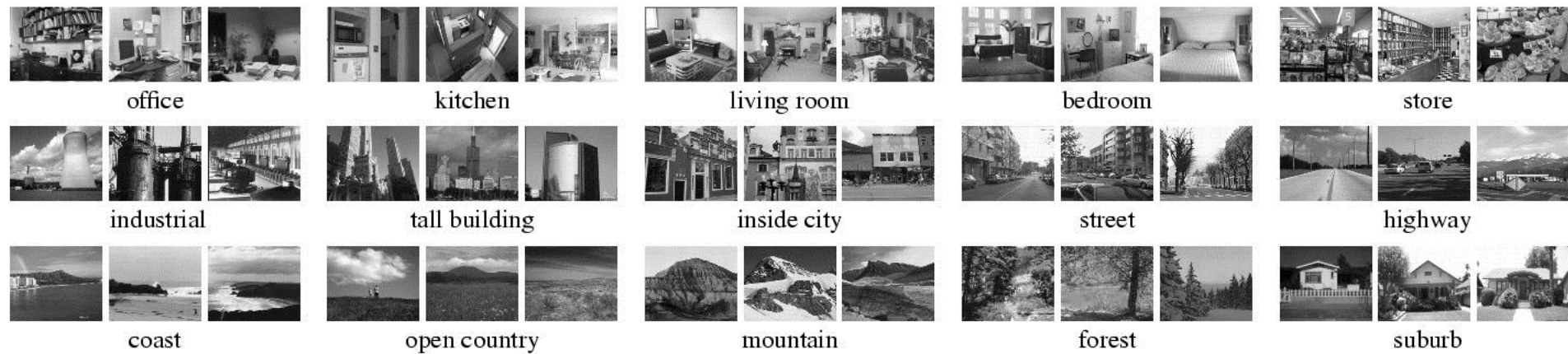


level 1



level 2

Scene category dataset

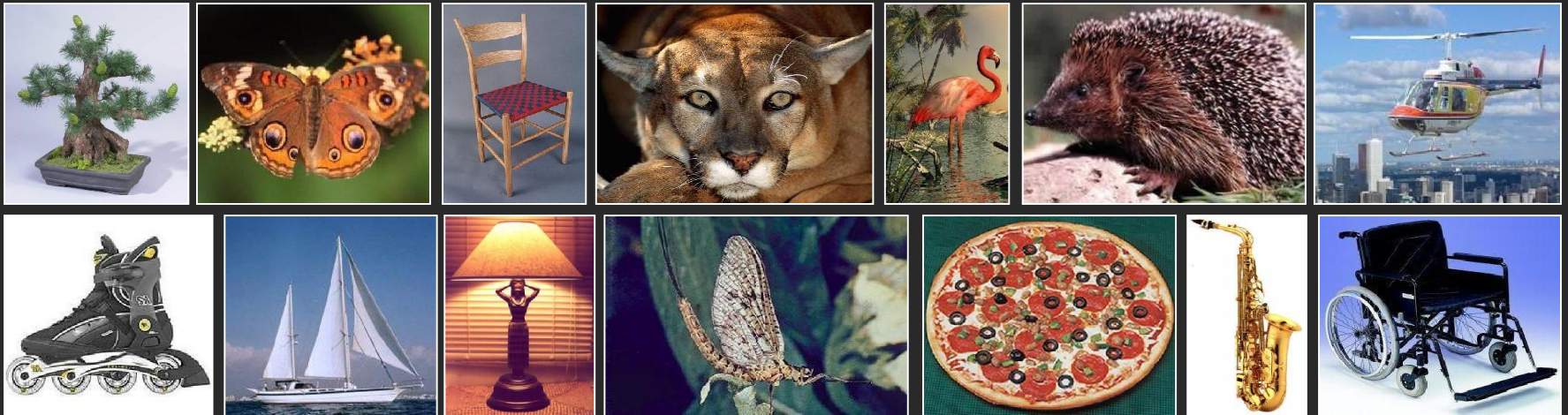


Multi-class classification results (100 training images per class)

Level	Weak features (vocabulary size: 16)		Strong features (vocabulary size: 200)	
	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

http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

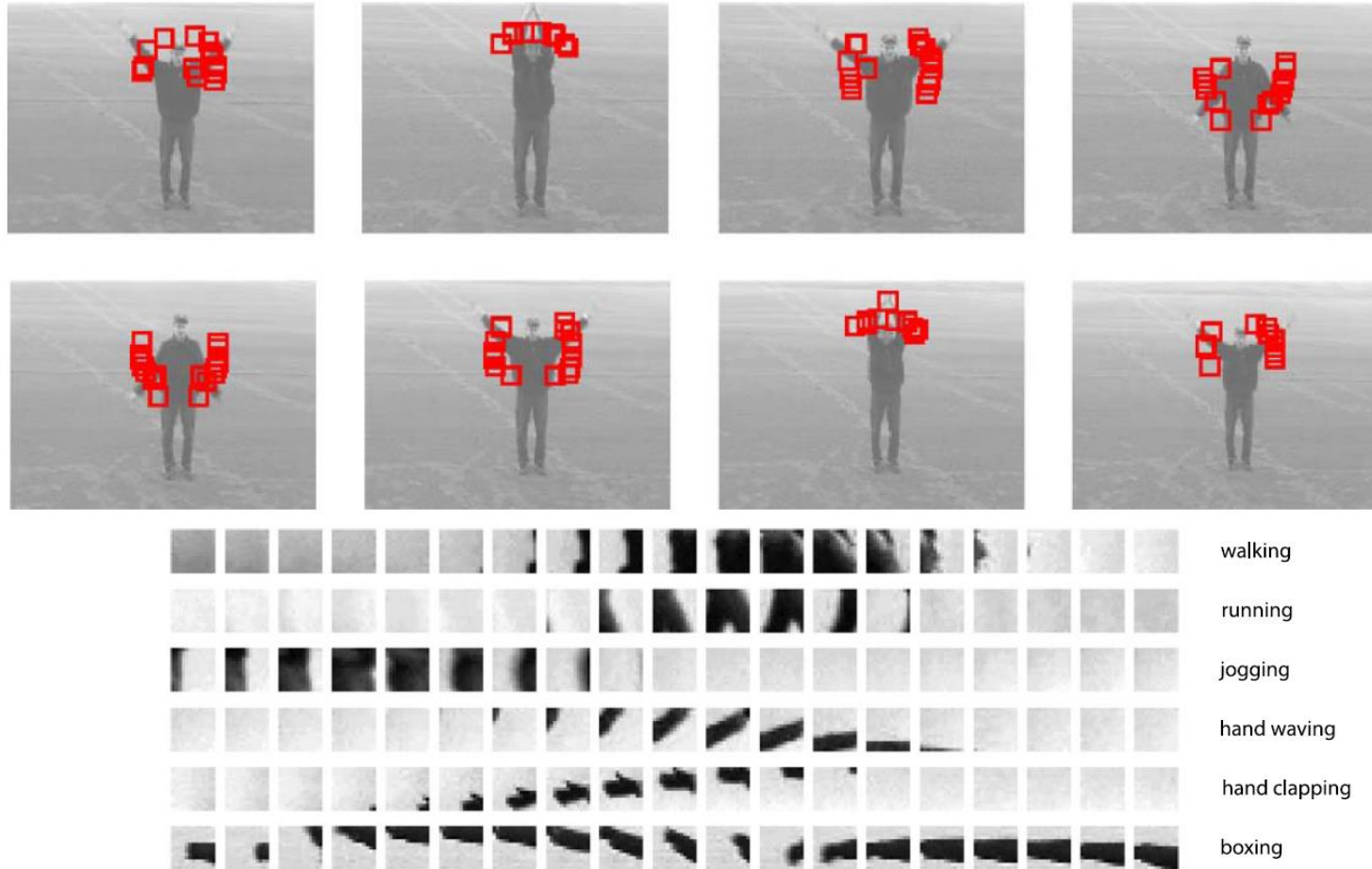


Multi-class classification results (30 training images per class)

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