Recognition

Lu Peng
School of Computer Science,
Beijing University of Posts and Telecommunications

Machine Vision Technology							
Semantic information				Metric 3D information			
Pixels	Segments	Images	Videos	Camera		Multi-view Geometry	
Convolutions Edges & Fitting Local features Texture	Segmentation Clustering	Recognition Detection	Motion Tracking	Camera Model	Camera Calibration	Epipolar Geometry	SFM
10	4	4	2	2	2	2	2

What we will learn today?

- Introduction to object recognition
  - Representation
  - Learning
  - Recognition
- Bag of Words models
  - Basic representation
  - Different learning and recognition algorithms

What are the different visual recognition tasks?

Source：Fei-Fei Li

Classification

Does this image contain a building? [yes/no]

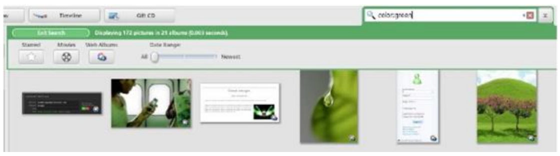
Source： Fei-Fei Li

Classification

Is this an beach?

Source： Fei-Fei Li

Image Search & Organizing photo collections



Source： Fei-Fei Li

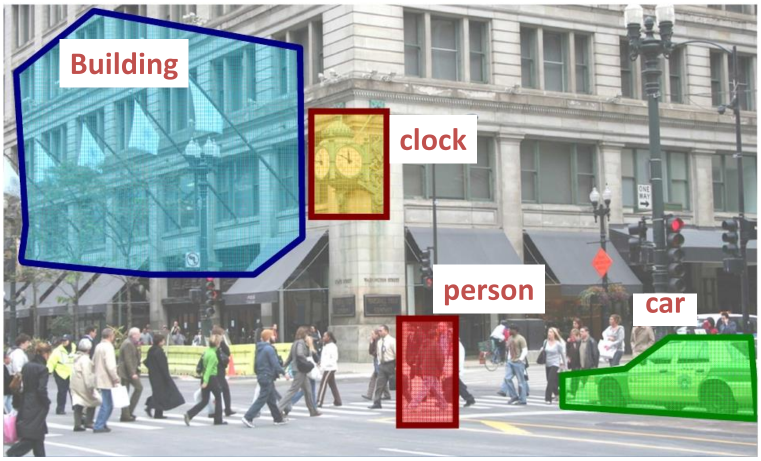
Detection

Does this image contain a car? [where?]

Source： Fei-Fei Li

Detection

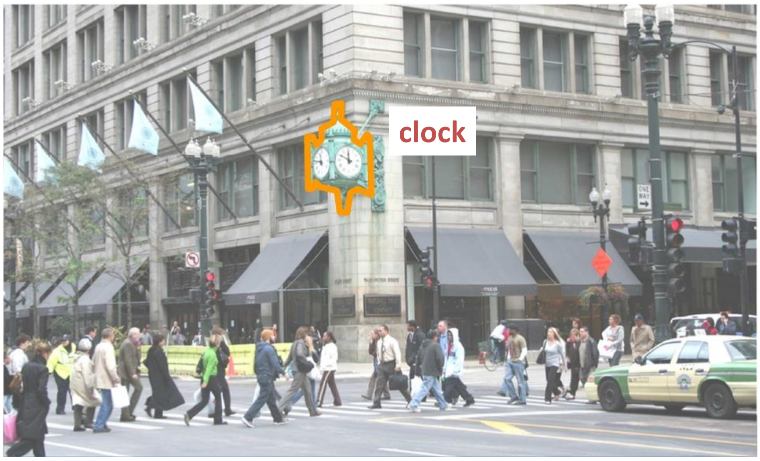
Which object does this image contain? [where?]



Source：Fei-Fei Li

Detection

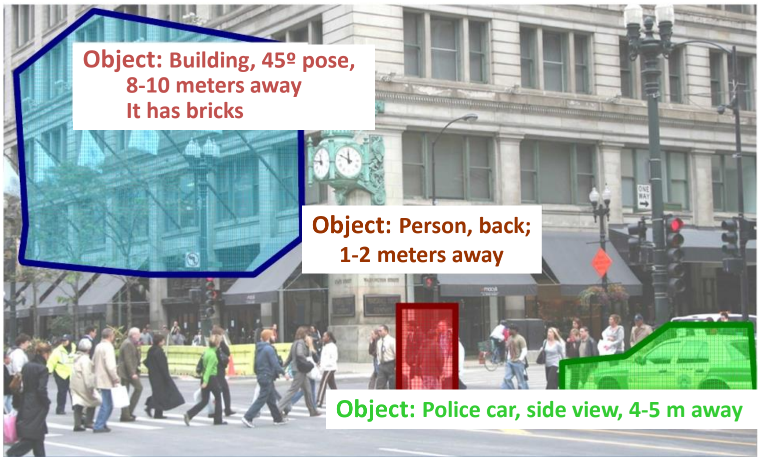
Accurate localization (segmentation)



Source：Fei-Fei Li

Detection

Estimating object semantic & geometric attributes



Source：Fei-Fei Li

Applications of computer vision

Computational photography

Assistive technologies

Surveillance

Security

Assistive driving

Source：Fei-Fei Li

Categorization vs Single instance recognition

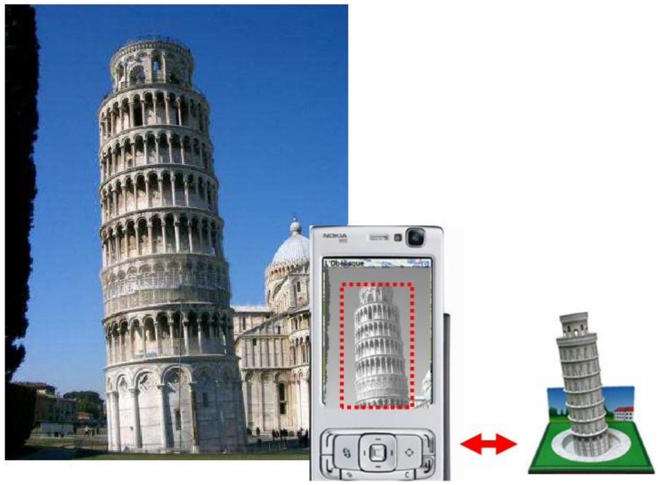
Does this image contain the Chicago Macy building's?

Categorization vs Single instance recognition

Where is the crunchy nut?

Source： Fei-Fei Li

Categorization vs Single instance recognition



Source： Fei-Fei Li

Activity or Event recognition

What are these people doing?

Source： Fei-Fei Li

Visual Recognition

- Design algorithms that are capable to
  - Classify images or videos
  - Detect and localize objects
  - Estimate semantic and geometrical attributes
  - Classify human activities and events

Why is this challenging?

Source： Fei-Fei Li

How many object categories are there?

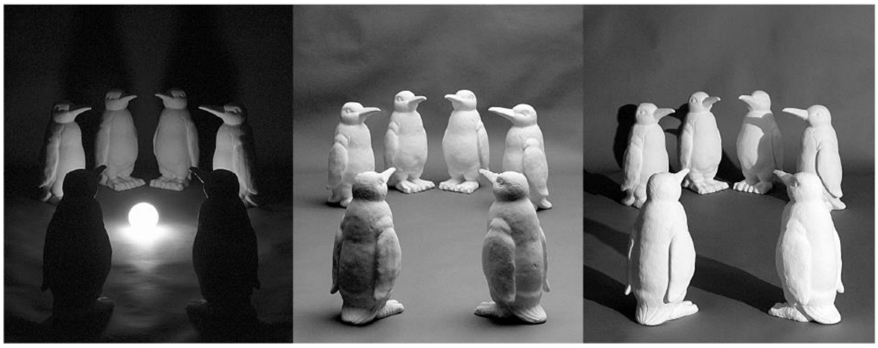
Source： Fei-Fei Li

Challenges: viewpoint variation

Michelangelo 1475-1564

Source： Fei-Fei Li

Challenges: illumination



Source： Fei-Fei Li

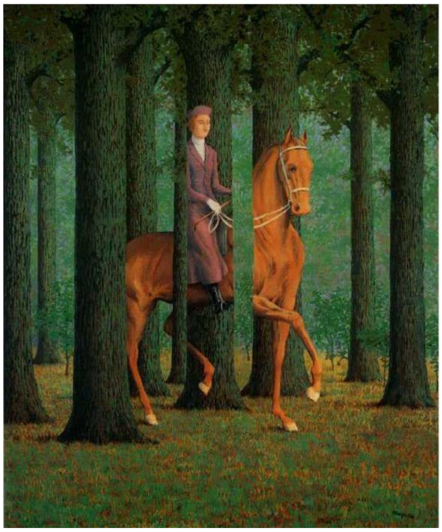
Challenges: scale

Source： Fei-Fei Li

Challenges: deformation

Source： Fei-Fei Li

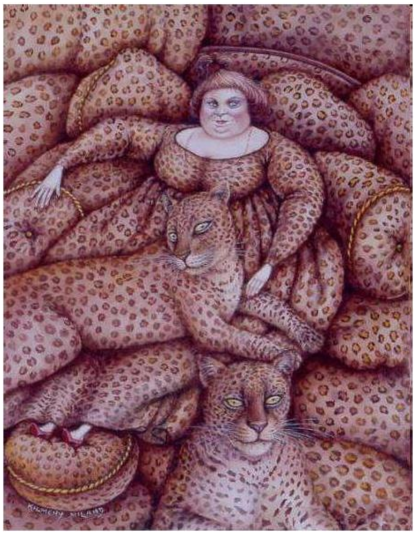
Challenges: occlusion



Magritte, 1957

Source：Fei-Fei Li

Challenges: background clutter



Kilmeny Niland. 1995

Source：Fei-Fei Li

Challenges: intra-class variation

Source：Fei-Fei Li

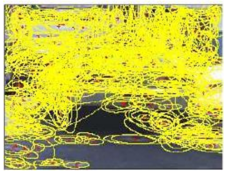
Basic issues

- Representation
  - How to represent an object category; which classification scheme?
- Learning
  - How to learn the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

Source：Fei-Fei Li

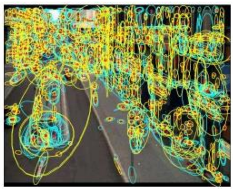
Representation

- Building blocks: Sampling strategies



Interest operators

Dense, uniformly



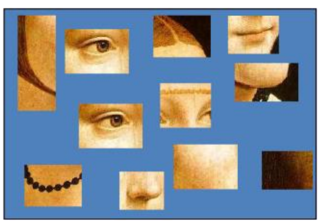
Multiple interest operators

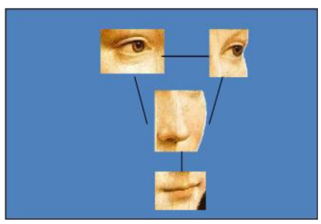
Randomly

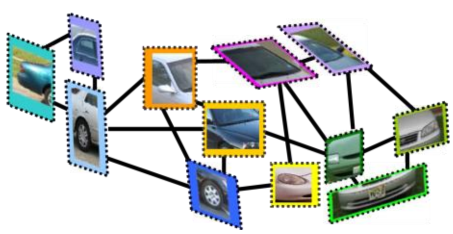
Source： Fei-Fei Li

Representation

– Appearance only or location and appearance





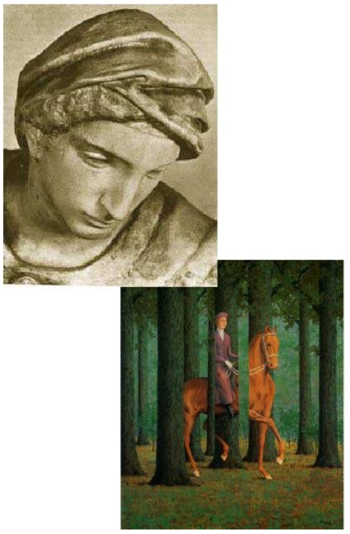


Source： Fei-Fei Li

Representation

Invariances

- View point
- Illumination
- Occlusion
- Scale
- Deformation
- Clutter
- etc.



Source： Fei-Fei Li

Representation

- To handle intra-class variability, it is convenient to describe an object categories using probabilistic models
- Object models: Generative vs Discriminative vs hybrid

Source： Fei-Fei Li

Object categorization: the statistical viewpoint

$p(\ zebra \mid image\ )$

vs.

$p(\ no\ zebra \mid image\ )$

- Bayes rule: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

Source： Fei-Fei Li

Object categorization: the statistical viewpoint

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

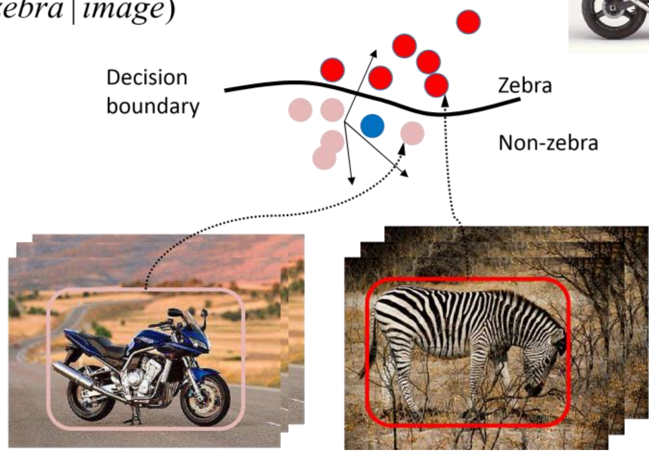
$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

Source： Fei-Fei Li

Discriminative models

- Modeling the posterior ratio:

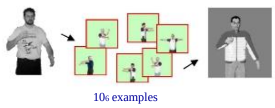
$$\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$$



Source： Fei-Fei Li

Discriminative models

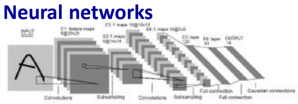
Nearest neighbor



10^6 examples

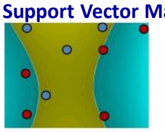
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

Neural networks



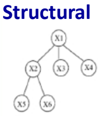
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines



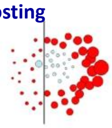
Guyon, Vapnik, Heisele,
Serre, Poggio…

Latent SVM
Structural SVM



Felzenszwalb 00
Ramanan 03...

Boosting



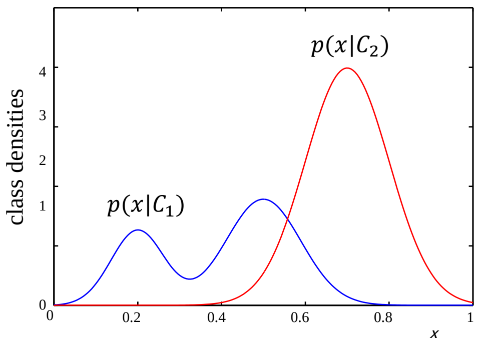
Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,…

Source: VittorioFerrari, Kristen Grauman, AntonioTorralba

Generative models

- Modeling the likelihood ratio:

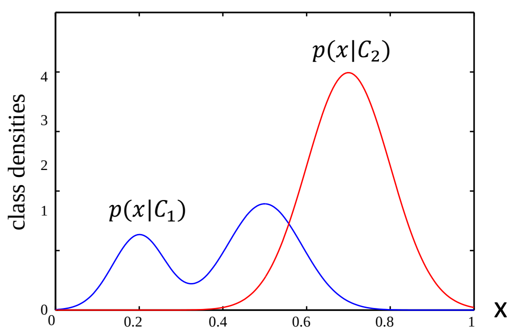
$$\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}$$



Source: Fei-Fei Li

Generative models

P(image | zebra)	P(image | no zebra)
High	Low
Low	High



Generative models

- **Naïve Bayes classifier**
  - ➢ Csurka Bray, Dance & Fan, 2004
- **Hierarchical Bayesian topic models (e.g. pLSAand LDA)**
  - ➢ Object categorization: Sivic et al. 2005, Sudderth et al. 2005
  - ➢ Natural scene categorization: Fei-Fei et al. 2005
- **2D Part based models**
  - ➢ Constellation models: Weber et al 2000; Fergus et al 200
  - ➢ Star models: ISM (Leibe et al 05)
- **3D part based models:**
  - ➢ multi-aspects: Sun, et al, 2009

Source： Fei-Fei Li

Basic issues

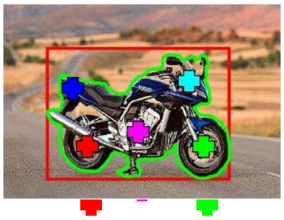
- Representation
  - – How to represent an object category; which classification scheme?
- Learning
  - – How to learn the classifier, given training data
- Recognition
  - – How the classifier is to be used on novel data

Source： Fei-Fei Li

Learning

- **Learning parameters: What are you maximizing?**

  Likelihood (Gen.) or performances on train/validation set (Disc.)
- **Level of supervision**
  - ➢ Manual segmentation; bounding box; image labels; noisy labels
- **Batch/incremental**
- **Priors**
- **Training images:**
  - ➢ Issue of overfitting
  - ➢ Negative images for discriminative methods



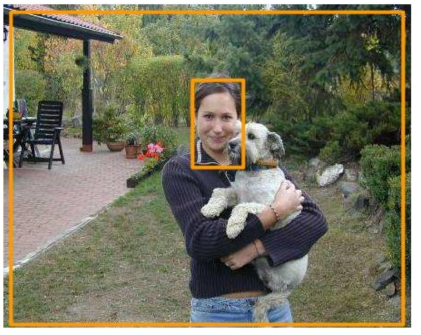
Source： Fei-Fei Li

Basic issues

- Representation
  - How to represent an object category; which classification scheme?
- Learning
  - How to learn the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

Recognition

- Recognition task: classification, detection, etc. .



Source： Fei-Fei Li

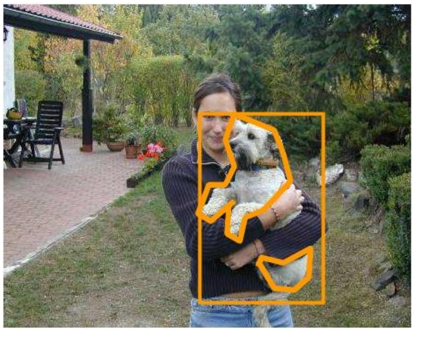
Recognition

- Recognition task: classification, detection, etc. .
- Search strategy: Sliding Windows
  - ➢ Simple
  - ➢ Computational complexity (x, y, S, θ, N of classes)
    - BSW by Lampert et al 08
    - Also, Alexe, et al 10

Source： Fei-Fei Li

Recognition

- Recognition task: classification, detection, etc. .
- Search strategy: Sliding Windows
  - ➢ Simple
  - ➢ Computational complexity (x, y, S, θ, N of classes)
    - BSW by Lampert et al 08
    - Also, Alexe, et al 10
  - ➢ Localization
    - Objects are not boxes
    - Prone to false positive



Source： Fei-Fei Li

Recognition

- Recognition task: classification, detection, etc. .
- Search strategy: Sliding Windows
  - ➢ Simple
  - ➢ Computational complexity (x, y, S, θ, N of classes)
    - BSW by Lampert et al 08
    - Also, Alexe, et al 10
  - ➢ Localization
    - Objects are not boxes
    - Prone to false positive

Non max suppression:
Canny '86
....
Desai et al , 2009

Source： Fei-Fei Li

Basic issues

- Representation
  - How to represent an object category; which classification scheme?
- Learning
  - How to learn the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

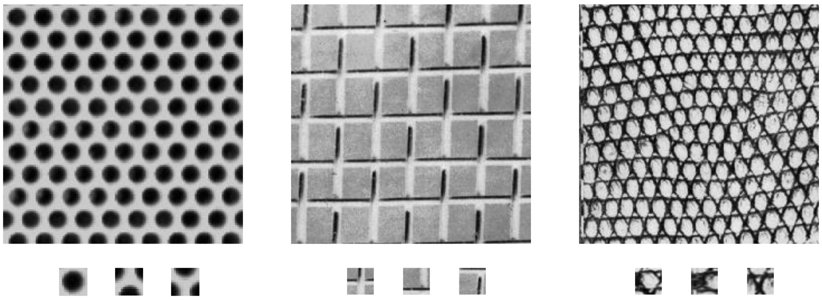
Bag-of-features models

Source: Lazebnik

Origin 1: Texture recognition

Texture is characterized by the repetition of basic elements or *textons*

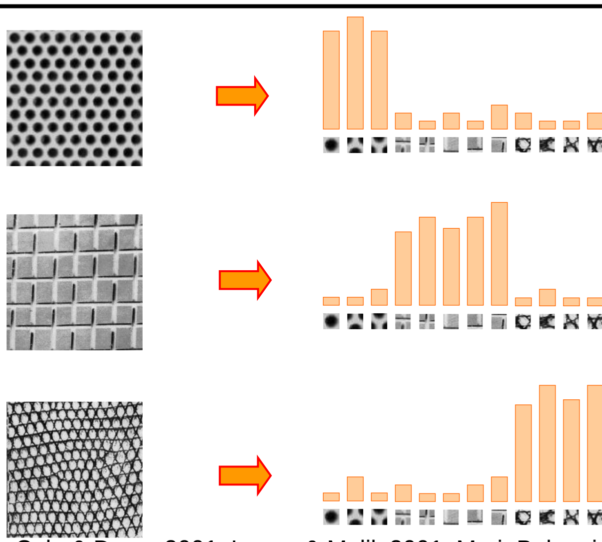
For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters



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

Source: Lazebnik

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

Source: Lazebnik

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)

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

Source: Lazebnik

Origin 2: 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/`

Source: Lazebnik

Origin 2: Bag-of-words models

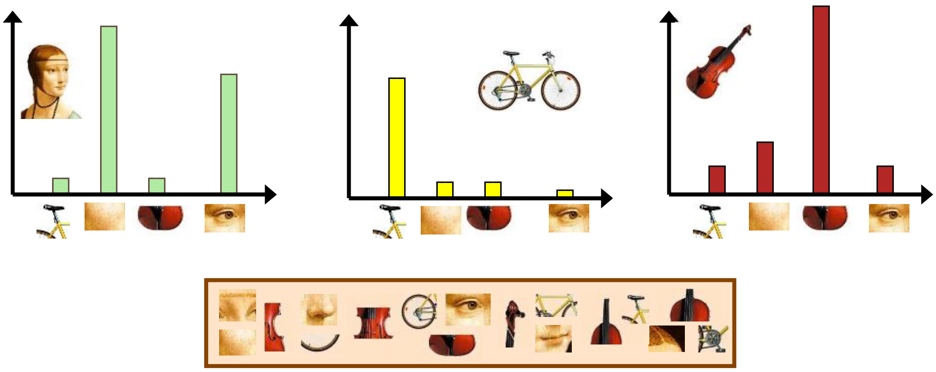
Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

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

Source: Lazebnik

Bag-of-features steps

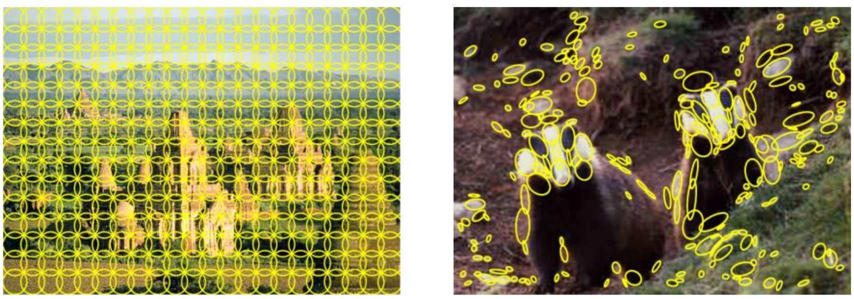
1. Extract features
2. Learn “visual vocabulary”
3. Represent images by frequencies of “visual words”



Source: Lazebnik

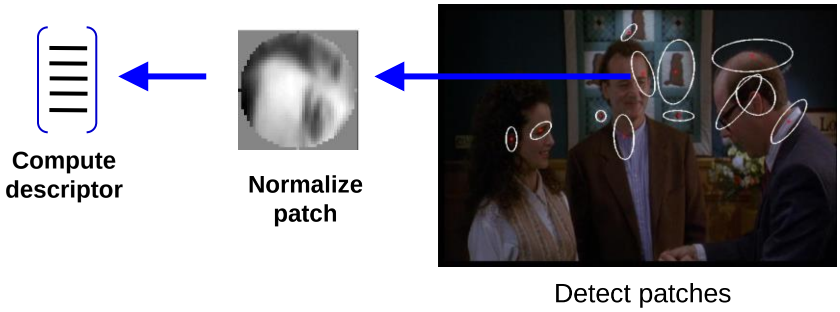
1. Feature extraction

Regular grid or interest regions



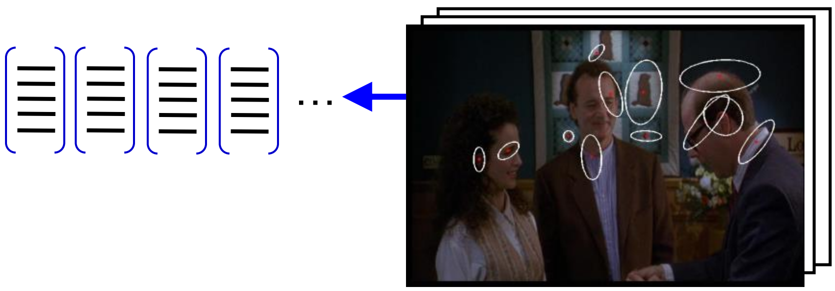
Source: Lazebnik

1. Feature extraction



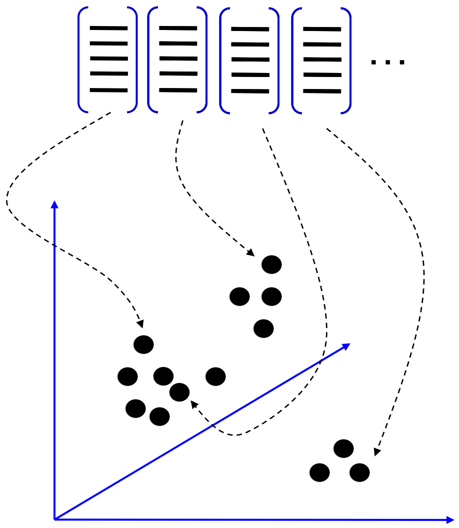
Source: Josef Sivic

1. Feature extraction



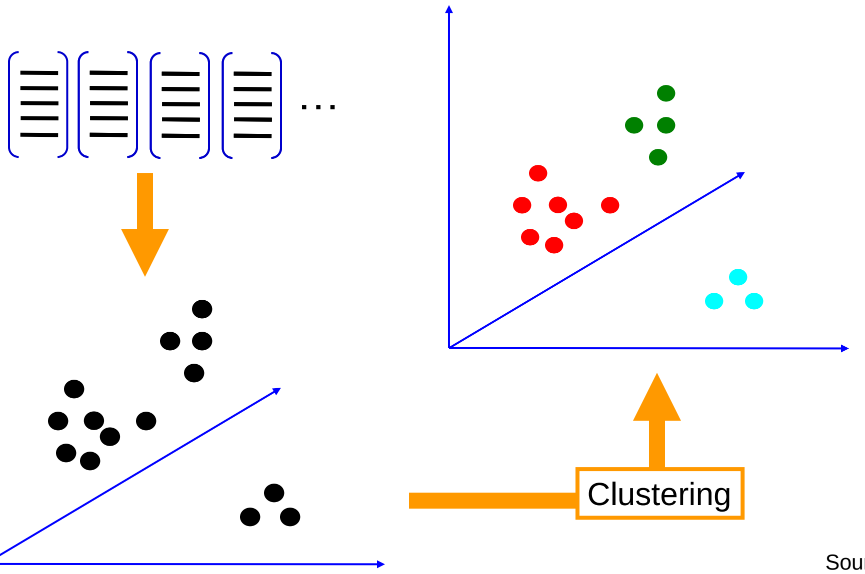
Source: Josef Sivic

2. Learning the visual vocabulary



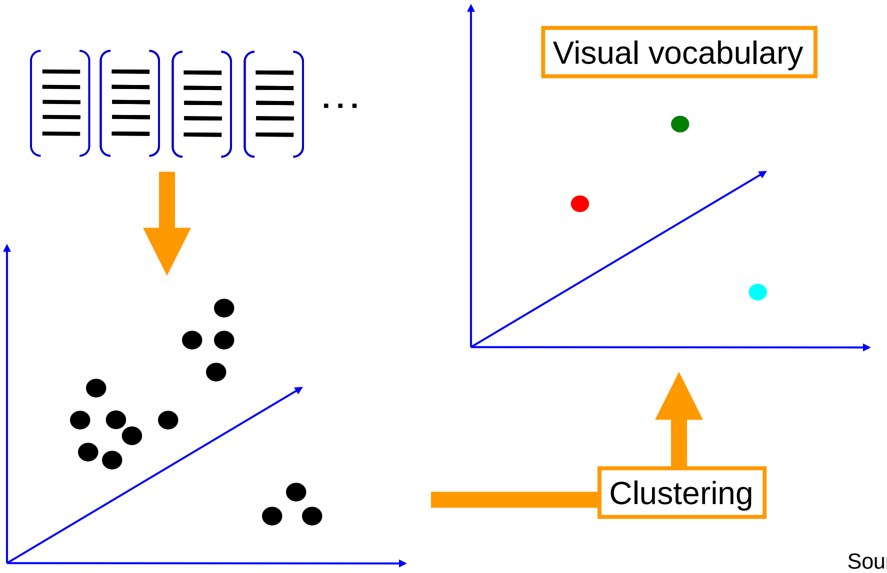
Source: Josef Sivic

2. Learning the visual vocabulary



Source: Josef Sivic

2. Learning the visual vocabulary



Source: 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

Source: Lazebnik

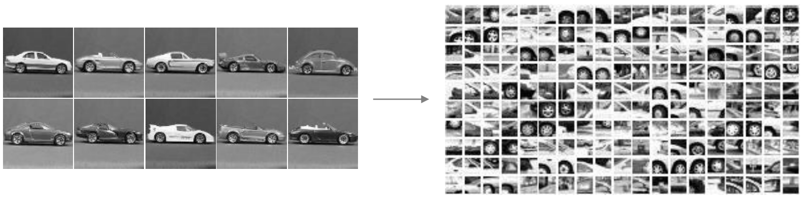
Clustering and vector quantization

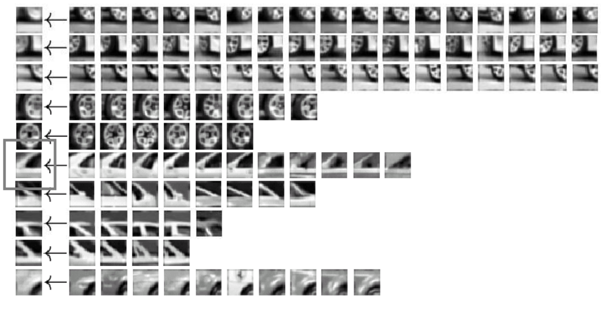
- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process
  - Each cluster center produced by k-means becomes a codevector
  - Codebook can be learned on separate training set
  - Provided the training set is sufficiently representative, the codebook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

Source: Lazebnik

Example codebook





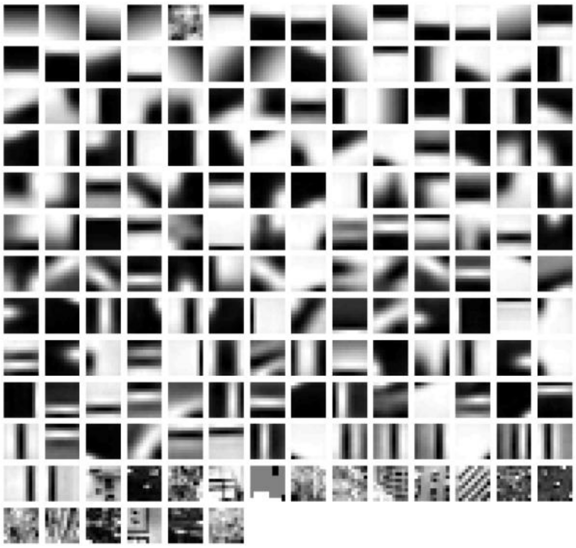
Appearance codebook

Source: B. Leibe

Another codebook

Source: B. Leibe

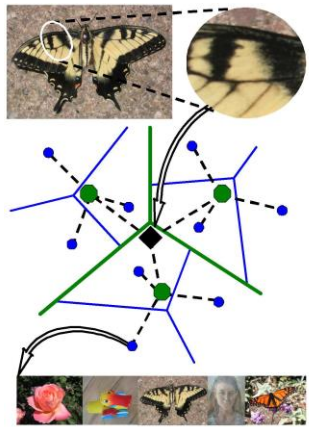
Yet another codebook



Source: Fei-Fei Li

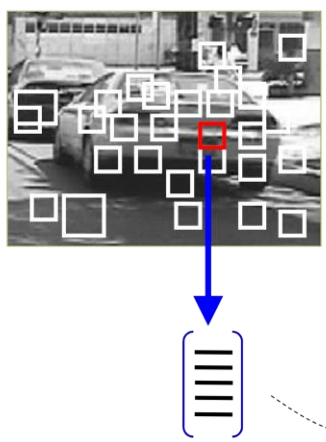
Visual vocabularies: Issues

- 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)

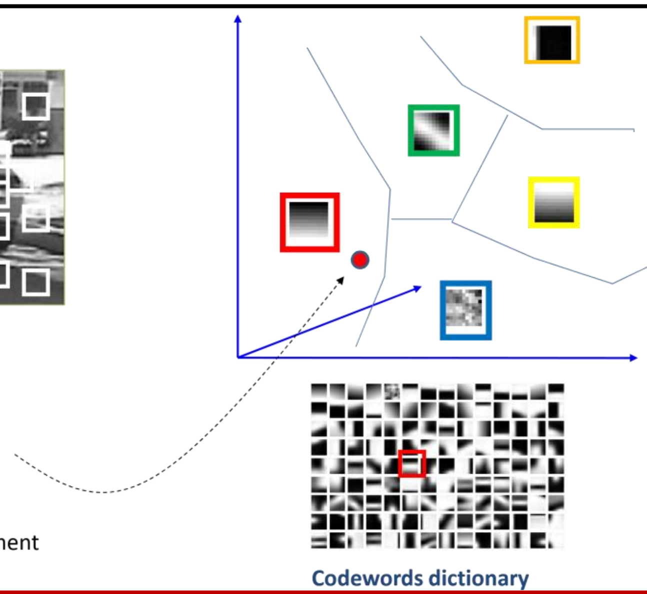


Source: Lazebnik

3. Bag of word representation

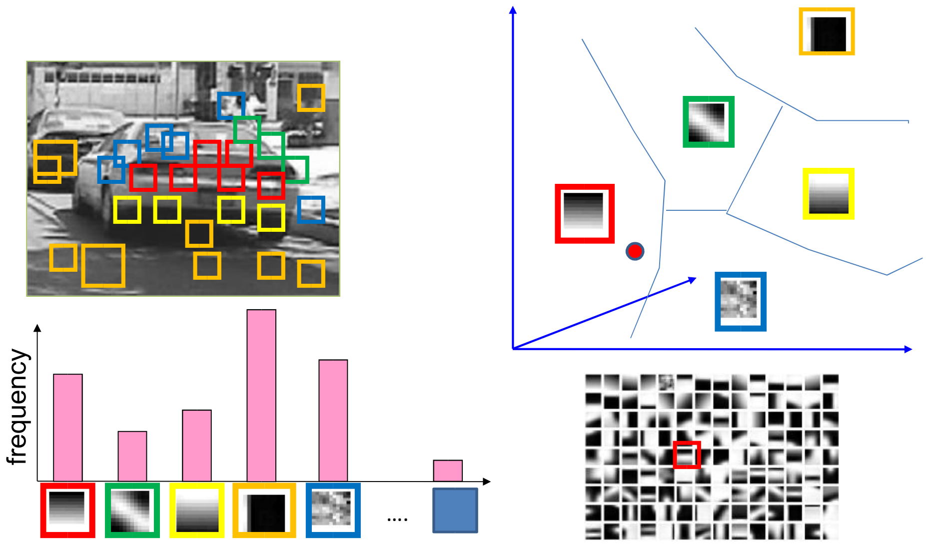


- Nearest neighbors assignment
- K-D tree search strategy



Source： Fei-Fei Li

3. Bag of word representation



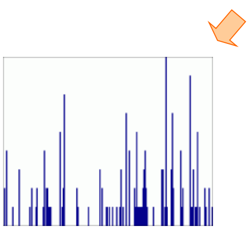
Source： Fei-Fei Li

Spatial pyramid representation

Lazebnik, Schmid & Ponce (CVPR 2006)

Extension of a bag of features

Locally orderless representation at several levels of resolution



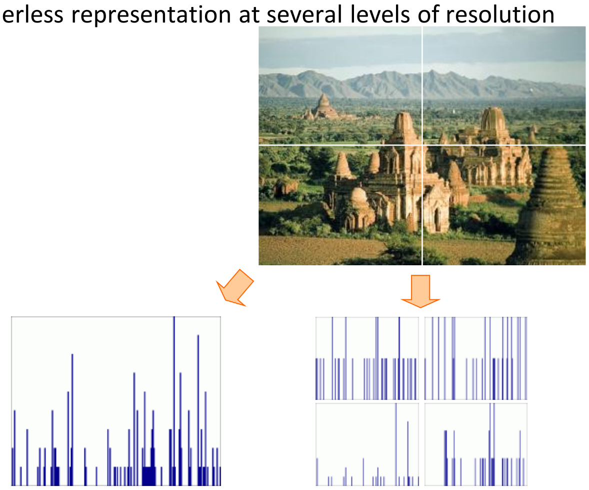
level 0

Spatial pyramid representation

Lazebnik, Schmid & Ponce (CVPR 2006)

Extension of a bag of features

Locally orderless representation at several levels of resolution



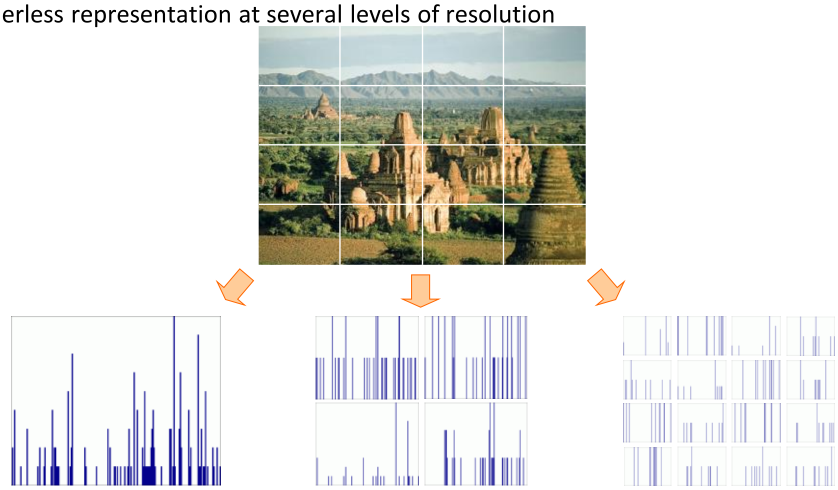
Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

Lazebnik, Schmid & Ponce (CVPR 2006)

Extension of a bag of features

Locally orderless representation at several levels of resolution



Scene category dataset

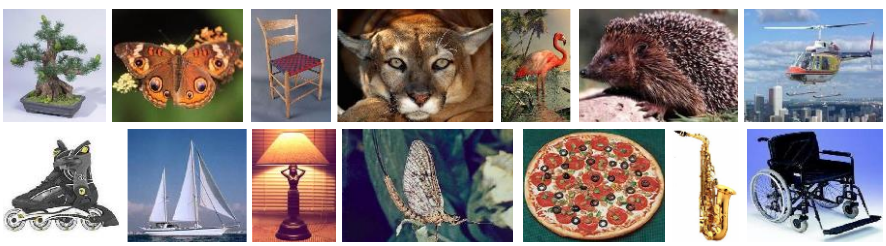
Lazebnik, Schmid & Ponce (CVPR 2006)

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 × 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

Lazebnik, Schmid & Ponce (CVPR 2006)



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

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