Course: Computer Vision

Unit 4: Object recognition

Introduction to Object Recognition

Luis Baumela

Universidad Politécnica de Madrid

Object Recognition

0. Computer vision and object recognition

1. Object recognition in perspective
    - Historical approach
    - Shallow approach
    - Deep approach

2. Fundamentals of object recognition. The shallow approach
    - How, where to describe?
    - Mid level representations
    - Pooling

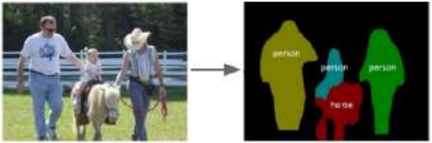
Computer Vision

- What is computer vision?

Related problems:

- Reconstruccion
- Segmentation
- Tracking
- **Recognition**

Instance segmentation



Object detection

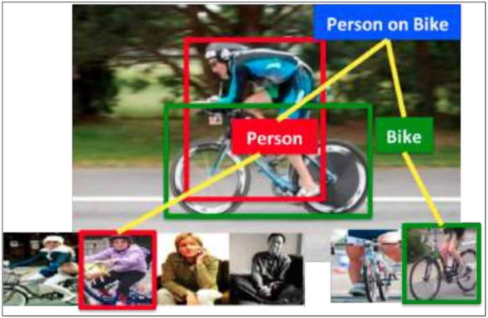


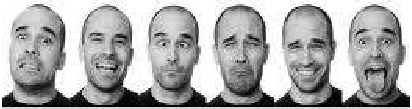
Image classification

container ship

Computer Vision

- Why is object recognition hard?
  - High variability of natural object classes
  - Lighting contrast, shadows, specularities
  - Geometric variability
  - Clutter, occlusion
  - Context
  - Deformation







Object Recognition

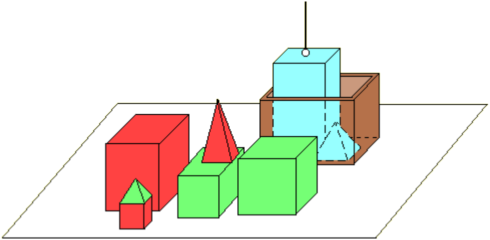
0. Computer vision and object recognition

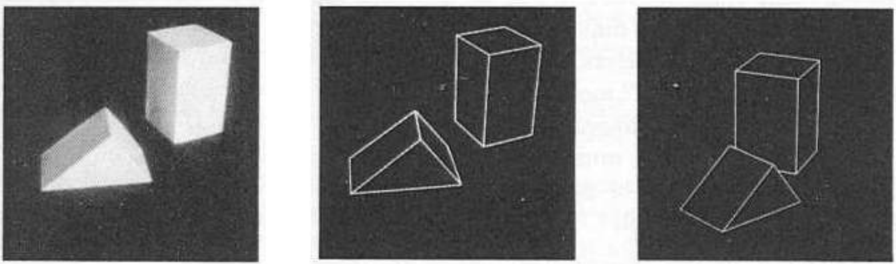
1. Object recognition in perspective
- **Historical approach**
- **Shallow approach**
- **Deep approach**

2. Fundamentals of object recognition. The shallow approach
- How, where to describe?
- Mid level representations
- Pooling

Object recognition in perspective

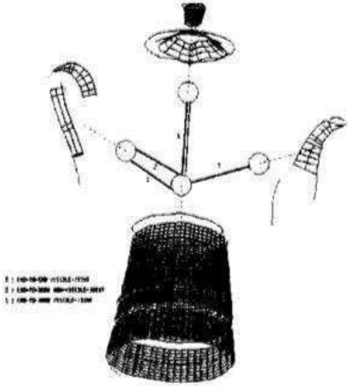
- Recognizing the blocks world (Roberts, 65)





Object recognition in perspective

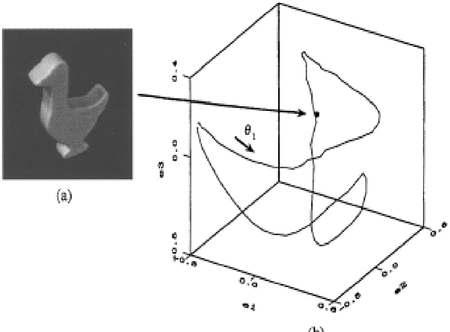
- Generalized cilinder representation (Binford, 71)



Object recognition in perspective

- Appearance manifolds (Murase, 95)

An image is represented by an n-dimensional vector.
Objects are represented as low dimensional manifolds embedded in n-dimensional space.

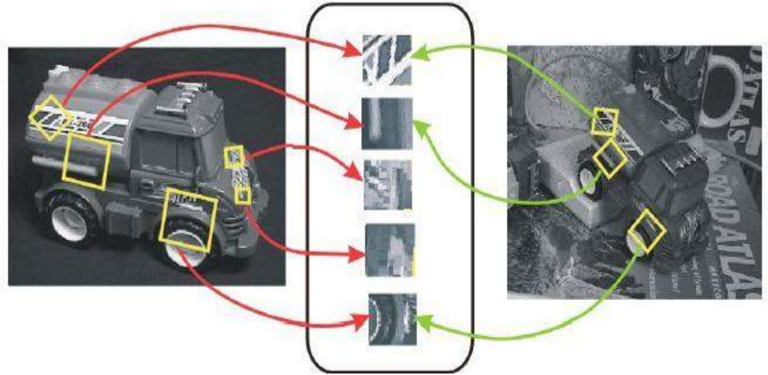


(a) (b)

Object recognition in perspective

- Local models of appearance (Lowe, 1999)

Image content represented by local features invariant to translation, rotation, scale and other imaging parameters.

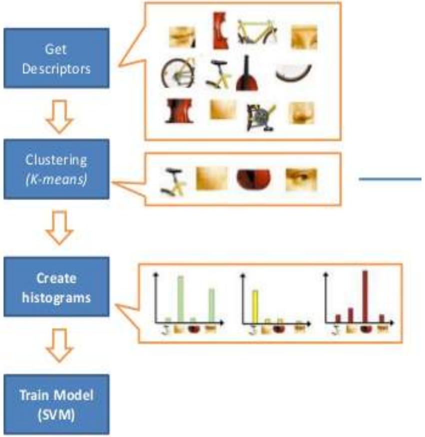


Object recognition in perspective

- Mid level representations (Sivic 2003, Lazebnik 2006)

Image content represented by the agreggation of local features into mid-level representations, before classification.

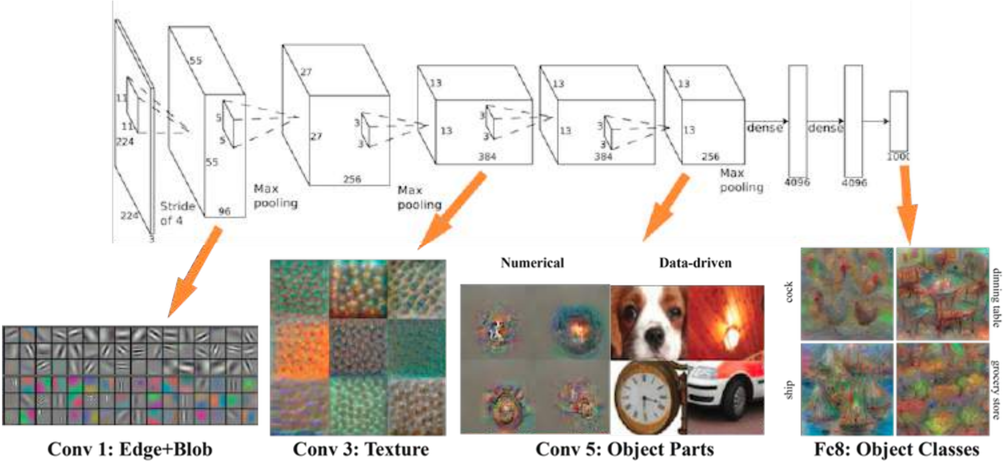
1. Low-level features
2. Mid-level representation
3. Pooling/aggregation
4. Classification



Object recognition in perspective

- Trainable hierarchical representations (Krizhevsky, 2012)

  Image content represented by the agreggation of local features into a hierarchy of representations AUTOMATICALLY trained.



http://vision03.csail.mit.edu/cnn_art/index.html

Object recognition in perspective

- Object recognition challenges

  Caltech 101 (2004), Caltech 256 (2007), Pascal VOC (2006-2012) , .

- Image Large Scale Visual Recognition Challenge

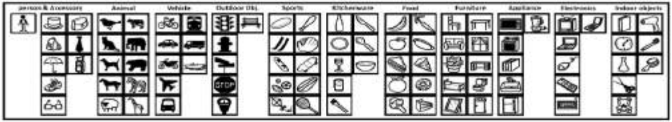
  (Russakovsky, 2015)

  - 1000 categories
  - 1.4 M images
  - ~ 3 instances per image

IMAGENET

- 15.000 visual categories
- 10 M labeled images
- ~ 700 images/category

- Common Objects in Context (Microsoft) (Li, 2014)
  - 91 categories, 328 k images
  - 2.5 M instances (~ 7.7 per image)
  - Every instance fully segmented



Object recognition in perspective

- Image Large Scale Visual Recognition Challenge

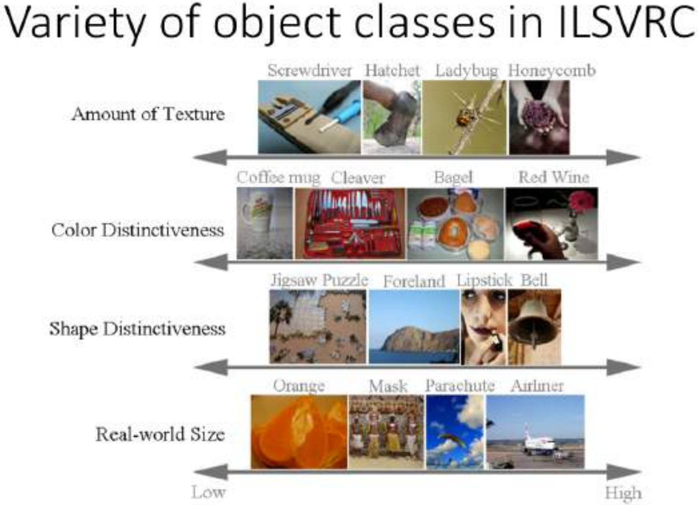


image-net.org

Object recognition in perspective

- Image Large Scale Visual Recognition Challenge

image-net.org

Object recognition in perspective

- Image Large Scale Visual Recognition Challenge

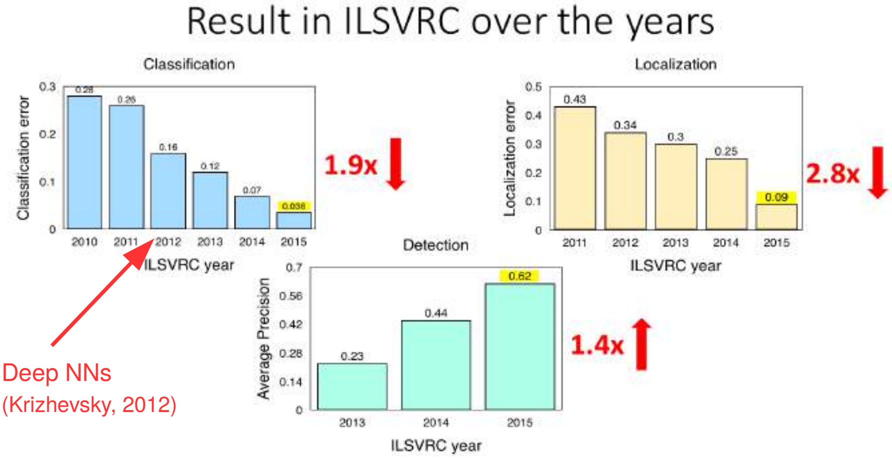


image-net.org

Object Recognition

0. Computer vision and object recognition

1. Object recognition in perspective
    - Historical approach
    - Shallow approach
    - Deep approach

2. Fundamentals of object recognition. The shallow approach
    - **How, where to describe?**
    - **Mid level representations**
    - **Pooling**

How to describe?

- Problem statement

We want to recognize objects inspite of the large variability of object clases, changes in appearance caused by illumination, geometry, deformation, etc.

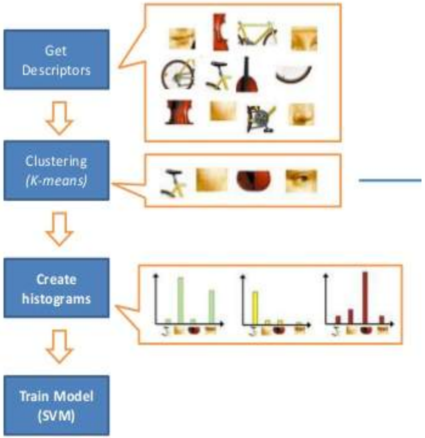
An appropriate image description invariant to most of those variations will be the key to success.

How to describe?

- Problem statement

Image content represented by the agreggation of local features into mid-level representations, before classification.

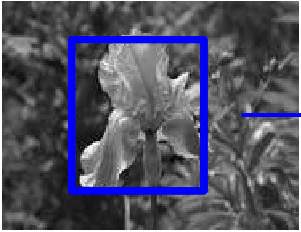
1. Low-level features
2. Mid-level representation
3. Pooling/aggregation
4. Classification



How to describe?

- Appearance-based descriptions

Describe the image using a vector of image pixel values



45	60	98	127	132	133	137	133
46	65	98	123	126	128	131	133
47	65	96	115	119	123	135	137
47	63	91	107	113	122	138	134
50	59	80	97	110	123	133	134
49	53	68	83	97	113	128	133
50	50	58	70	84	102	116	126
50	50	52	58	69	86	101	120

Problem:

- Storage requirements
- Invariance

How to describe?

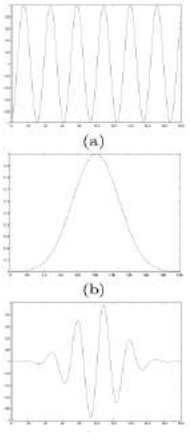
- Linear filters

  Describe the image using the responses of a set of filters.

  Gabor Filters:

  A Gaussian kernel function modulated by a sinusoidal plane wave

$$g_e(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \cos(2\pi\omega_0 x)$$



It responds to some frequency in a localized part of the signal.

The responses of simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions.

How to describe?

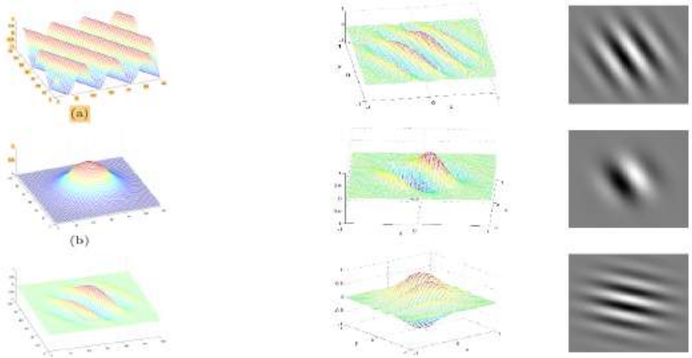
- Linear filters

  Describe the image using the responses of a set of filters.

  Gabor Filters:

  A Gaussian kernel function modulated by a sinusoidal 2D wave

$$g_e(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x}+\frac{y^2}{\sigma_y})} \cos(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y)$$



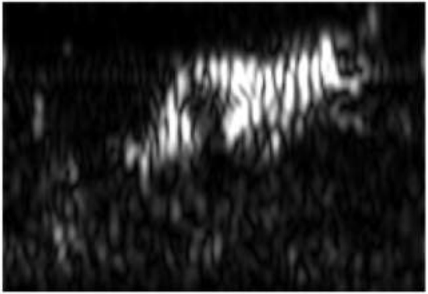
How to describe?

- Linear filters

  Describe the image using the responses of a set of filters.

  Gabor Filters:

  Sample Response



They have been used for recognizing facial expressions (Wu, 2010).

How to describe?

- Gradient information

Describe the image using information from image gradients.

Object appearance may be characterized by computing differences between sum of pixels in rectangles.

- Haar-like waveletts

Object appearance and shape can be charaterized by the distribution of local intensity gradients.

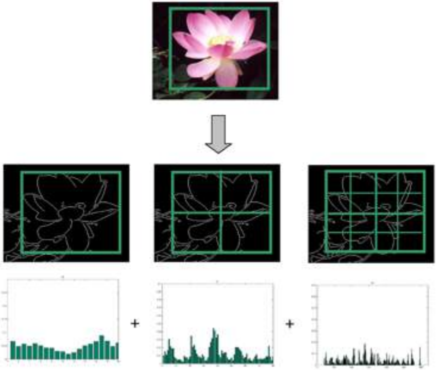
- Histograms of oriented gradients (HoG)
- Scale Invariant Feature Transform (SIFT)
- Speeded-Up Robust Features (SURF)

How to describe?

- Gradient information

  Describe the image using information from image gradients.

  Object appearance and shape can be charaterized by the distribution of local intensity gradients.

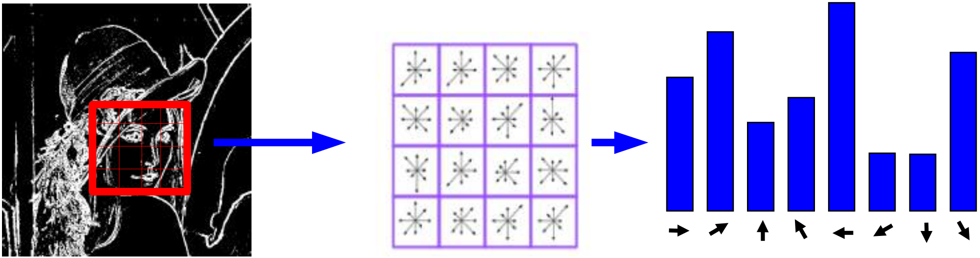


How to describe?

- Gradient information

Describe the image using information from image gradients.

Scale Invariant Feature Transform (SIFT):



Append 16 gradient histograms 8 bins each, 128 dimensional despriptor.

Lowe, IJCV, 2004.

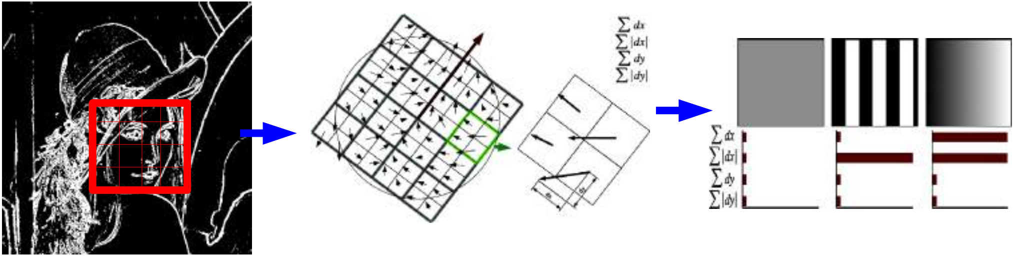
How to describe?

- Gradient information

  Describe the image using information from image gradients.

  Speeded-Up Robust Features (SURF):

  - Project a grid of size 4x4.
  - At each grid location compute the gradient of 5x5 regularly distributed points
  - Accumulate the values of $(\sum dx, \sum |dx|, \sum dy, \sum |dy|)$ in each grid point.
  - Append 16 histograms to form a 64 elements vector



Where to describe?

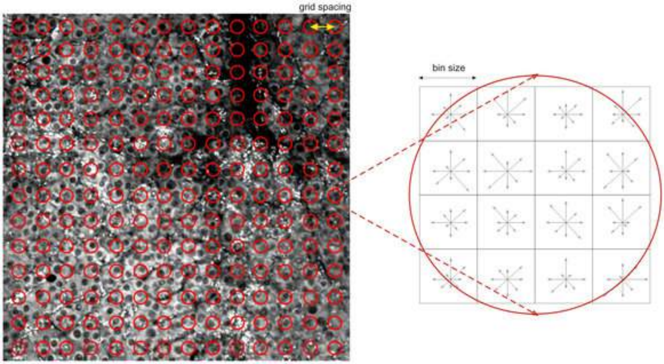
We can describe an image with the previous descriptors in various ways:

- Globally. Treating the whole image as a single object.
- Densely. On a dense grid of image locations.
- Superpixels. On small regions in the image.
- Sparsely.
  - On a ramdom set of locations
  - On a salient set of points.

Where to describe?

We can describe an image with the previous descriptors in various ways:

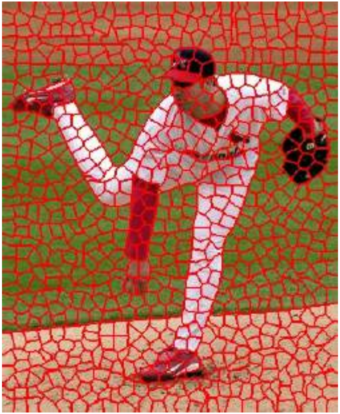
- Densely. On a dense grid of image locations.



Where to describe?

We can describe an image with the previous descriptors in various ways:

- Superpixels. On small regions in the image.



Superpixels, Ren, ICCV'03

Where to describe?

We can describe an image with the previous descriptors in various ways:

- Locally.
    - On a random set of locations
    - **On a salient set of points**.

Where to describe?

- Local representations

Represent the image or region as a set of descriptors over local image/region patches.

Locality reduces influence of:

- Partial occlusions & clutter
- Changes in illumination
- Some object deformations

But we lose

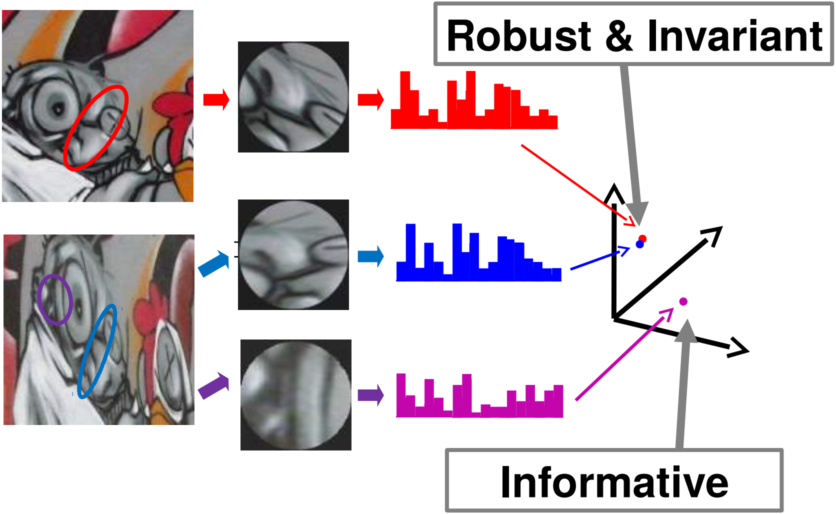
- Global image description

Local descriptions, “per se” are not invariant to

- Changes in orientation, deformation
- Chages in scale

Where to describe?

Local representations and **invariance**



Where to describe?

Local representations and **invariance** are the key concepts in the traditional object recognition paradigm.

We aim for a representation that is:

- Local: so robust to occlusion, clutter and ilumination.
- Invariant: approximatelly constant across rotations, scaling and some deformations.
- Robust: noise, blur, discretization do not change de representation.
- Informative: features can be matched to a large database of objects.
- Dense: many features can be generated even for small objects
- Accurate: precise location.
- Efficient: fast enough for the task at hand.

Where to describe?

- Interest operators

  We can achieve **scale, position, rotation** and **affine invariance** by adequatelly normalizing and describing patches that are local maxima of a second derivative operator in scale and position image spaces.

  Various operators combine these ideas to detect “interesting” or salient points in an image:

  - Hessian and Harris (Beudet'78, Harris'88)
  - Laplacian, DoG (Lindeberg'98, Lowe'99)
  - Harris-Hessian/Affine-Laplace (Mikolajczyk'01 y '04)
  - MSER (Matas'02)

  A comparison is given in (Mikolajczyk 2005).

Where to describe?

- Interest operators

  SIFT (Lowe, 2004) interest points are local maxima of Laplacian of Gaussian, which are efficiently approximated by a difference of gaussian.

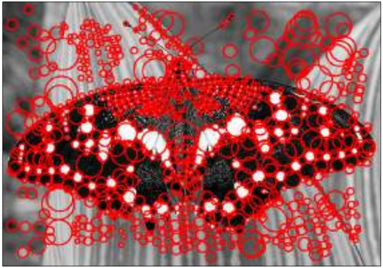
Where to describe?

- Interest operators

SURF (Bay, 2008) interest points are local maxima of scaled determinant of hessian (DoH)

$$\mathcal{H}(\sigma) = \sigma^4 (L_{xx}(\sigma) L_{yy}(\sigma) - L_{xy}(\sigma)^2)$$

which has extrema values at “blobs”



Local maxima give us the position and scale of blobs

The DoH is widely used for its robustness (fires less on edges than LoG) and because it is able to detect *saddle points*.

Object representation

- Bag of words

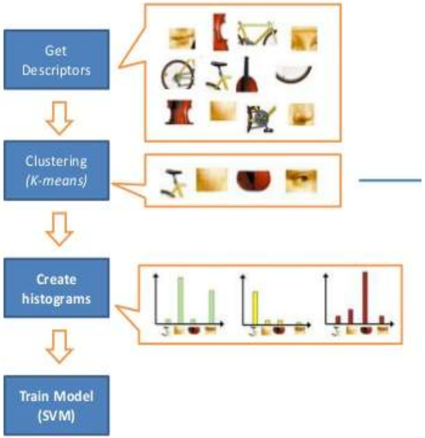
Group local features with similar appearance into a single object.

1. Low-level features

2. Mid-level representation

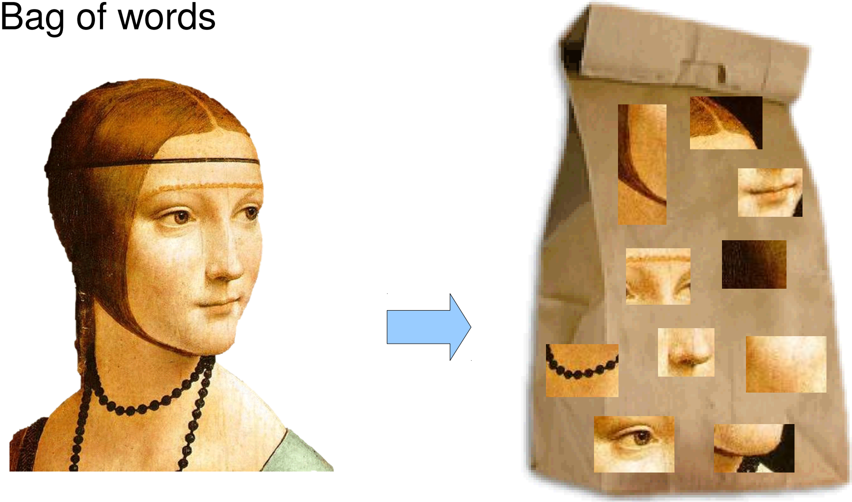
3. Pooling/aggregation

4. Classification



Object representation

- Bag of words



Object

Model

Object representation

- Bag of words algorithm

  1. Extract features

- What features?

  SIFT, SURF, MSER

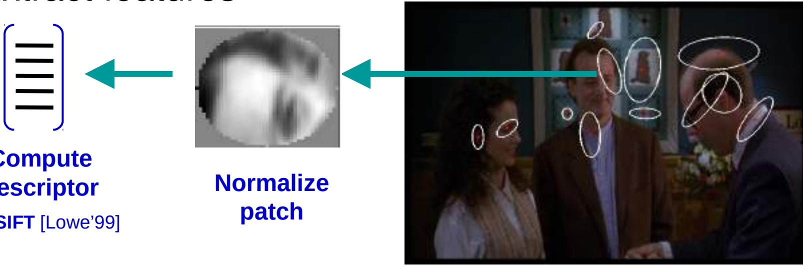
- Where?

  Densely, **using an interest point detector**, randomly

Object representation

- Bag of words algorithm

  1. Extract features



Compute descriptor

e.g. **SIFT** [Lowe'99]

Normalize patch

Detect patches

[Mikojaczyk and Schmid '02]

[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

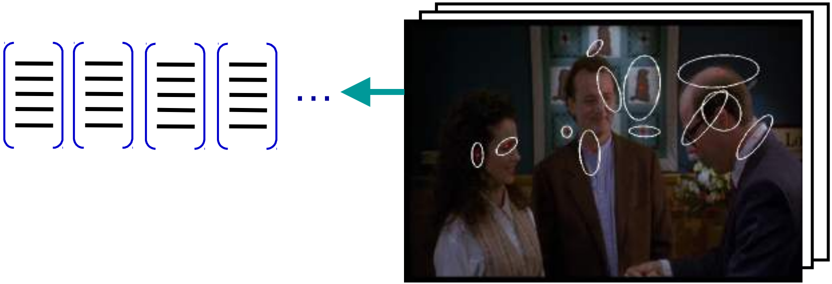
Local interest operator
or
Regular grid

Object representation

- Bag of words algorithm

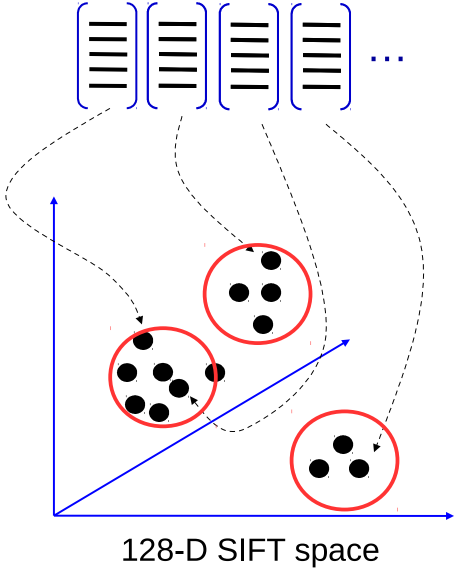
  1. Extract features

     For all images in the training set



Object representation

- Bag of words algoritm
  1. Extract features
  2. Learn “visual vocabulary”



Find clusters in the patch description space

Object representation

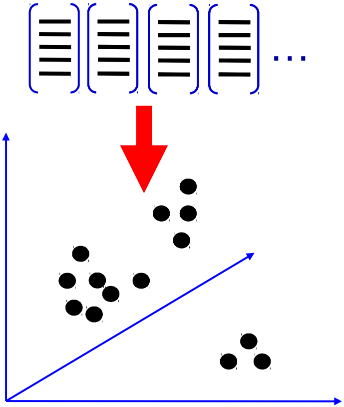
- Bag of words algoritm
  1. Extract features
  2. Learn “visual vocabulary”
  3. Quantize features

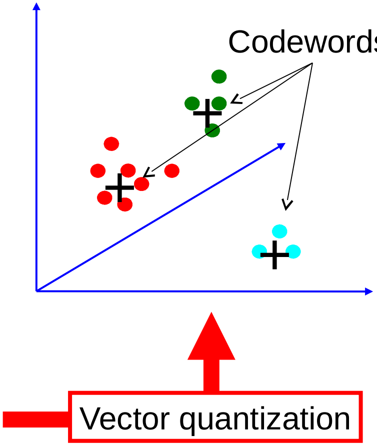
The codebook is used for quantizing features:

- 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

Object representation

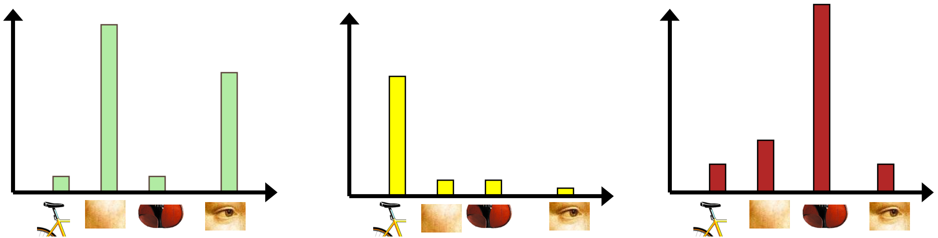
- Bag of words algoritm
  1. Extract features
  2. Learn “visual vocabulary”
  3. Quantize features





Object representation

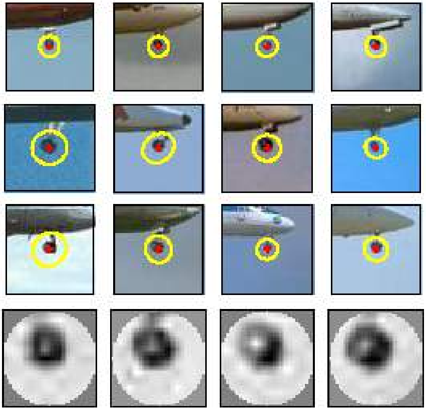
- Bag of words algoritm
  1. Extract features
  2. Learn “visual vocabulary”
  3. Quantize features
  4. Represent images with frequencies of visual words

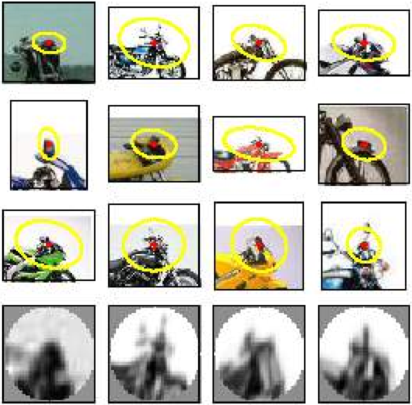


Object representation

- Bag of words

Sample visual words





Pooling spatial information

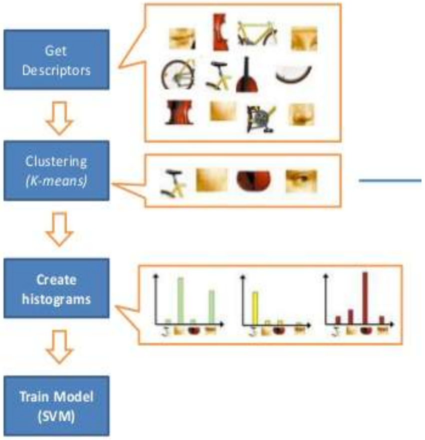
Group detections of visual words in different parts of the image

1. Low-level features

2. Mid-level representation

3. Pooling/aggregation

4. Classification

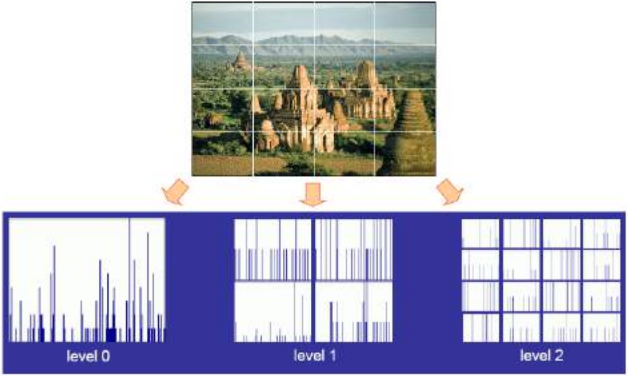


Pooling spatial information

- Generative models
- Discriminative models

  Directly estimate the image/object class from

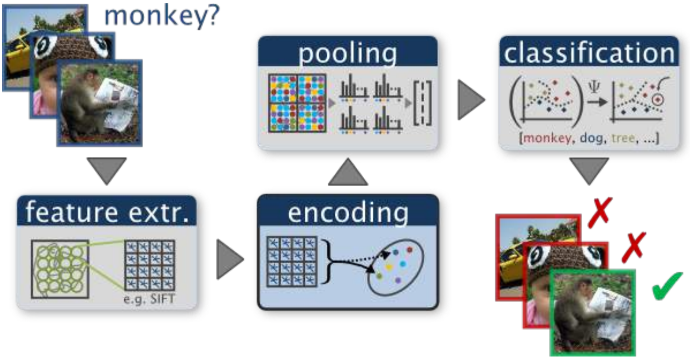
  Bag of word pyramid



Lazebnik, Schmid & Ponce, 2006

Standard object recognition approach

1. Extract low-level features
2. Compute mid-level representation (quantification)
3. Pool/aggregate spatial information
4. Classify



References

- T. O. Binford. Visual Perception by Computer. Proc. IEEE Conf. on Systems and Control, December 1971.
- H. Murase and S. Nayar. Learning and recognition of 3d objects from appearance. International Journal of Computer Vision, 14(1):5–24, 1995.
- L. G. Roberts. Machine perception of threedimensional solids. In Tippett, J. and Berkowitz, D. and Clapp, L. and Koester, C. and Vanderburgh, A., editor, Optical and Electrooptical Information processing, pages 159–197. MIT Press, 1965.
- D. G. Lowe. Object recognition from local scaleinvariant features. In Proceedings of the International Conference on Computer VisionVolume 2, 1999.
- J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In Int. Conf. on Conputer Vision (ICCV), 2003.
- S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2006.
- A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In Proc. Neural Information Processing Systems (NIPS), 2012.
- O. Russakovsky, et al. ImageNet Large Scale Visual Recognition Challenge. Int. Journla of Computer Vision, 115(3):211-252, 2015.
- T.Y. Li et al. Microsoft COCO: Common Objects in Context. Proc. European Conference on Computer Vision (ECCV), 2014.