Retrieval in multimedia

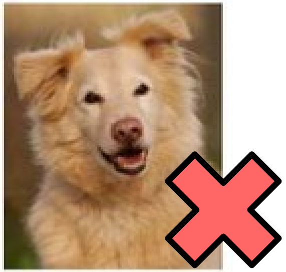
Image retrieval systems

- Images can be queried using
  - Metadata (text)
  - User annotations
  - Image features (content)
- Problems
  - Metadata is not complete/informative/available
  - User annotations not supported, unreliable

Images and text queries

- Images in web documents
  - Use text around image (URL element name, neighborhood)
  - Same principles as in text retrieval systems
- Example of searching for images with word »Sunset«

Sunset at Rocky Point



Frank Smiles
at Sunset

Sunset Beach

Images and text queries

Query: »tiger in woods«

Desired result

Obtained result

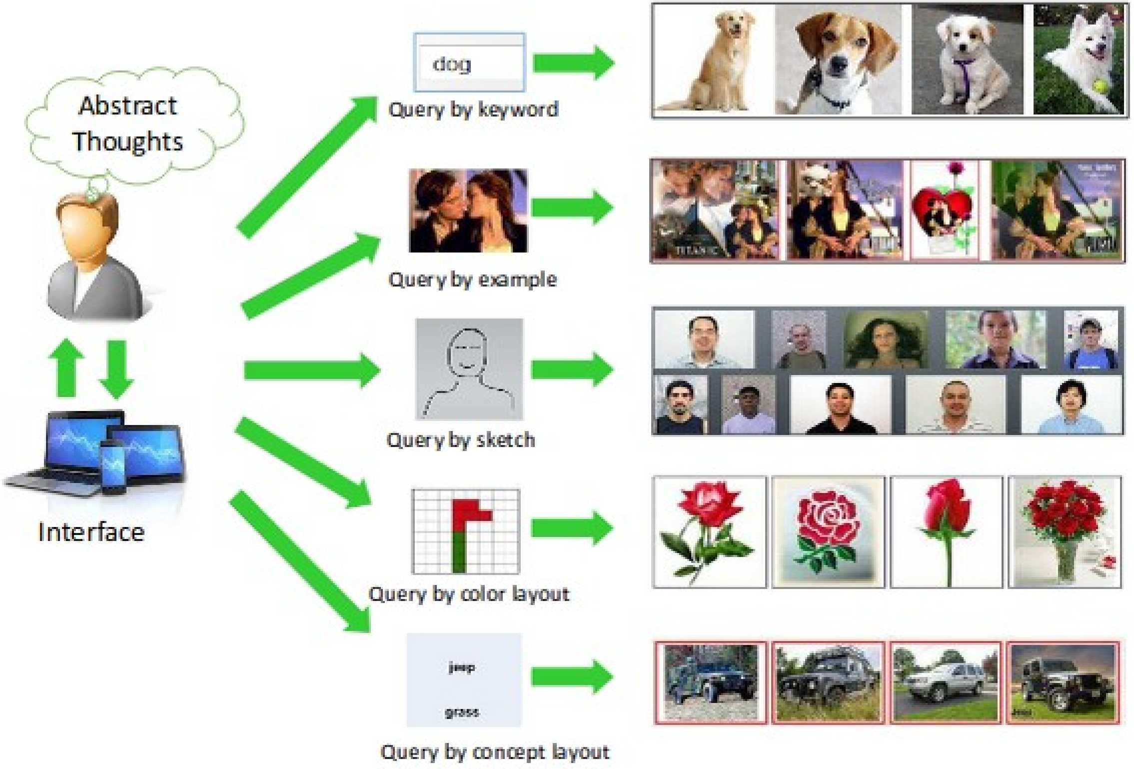
Problems with text queries

- Avoid using image content
  - Annotation bias
  - Metadata ambiguity
- Perceptual relevance
  - Impossible to describe composition
  - Abstract shapes

Development of retrieval systems that

encode image content directly

Image retrieval systems



Wengang Zhou, Houqiang Li, Qi Tian, Recent Advance in Content-based Image Retrieval: A Literature Survey, 2017

Querying image content

- Extract image content
  - Detecting object and categories
  - Describing relations, actions
  - Ambiguous problem
- Low-level features
  - Color
  - Texture
  - Shape
  - Structural elements

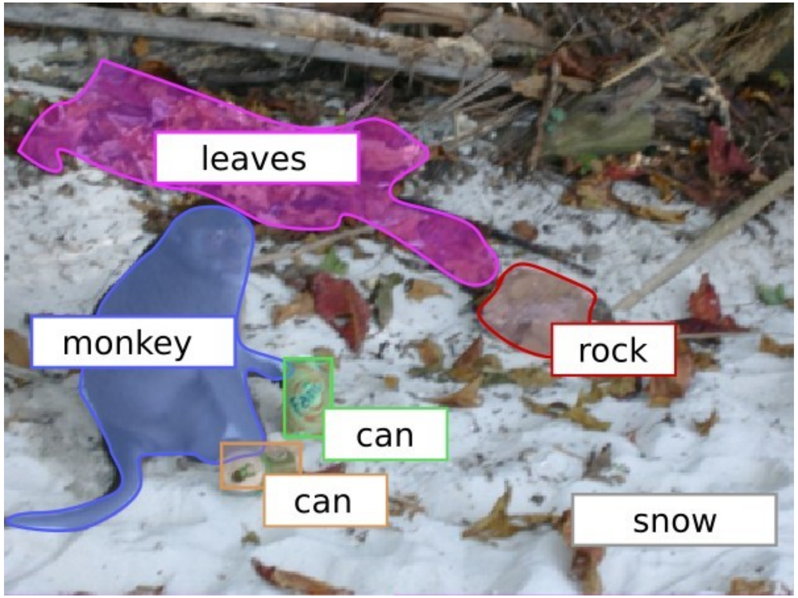


Image retrieval system

Querying by color

- Average color – no information about the distribution around average

μ_1 $\qquad d = \mu_1 - \mu_2 \qquad$ μ_2

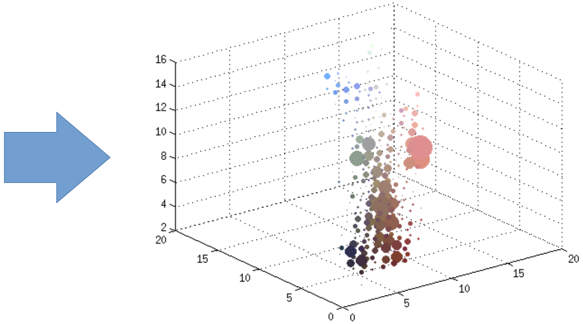
- Parametric distribution (Gaussian)

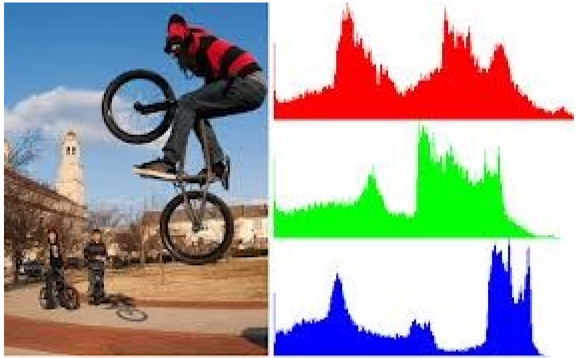
(μ_1, σ_1) $\qquad$ (μ_2, σ_2)

Bhattacharryya distance: $d = \frac{1}{8}(\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\left(\frac{|\boldsymbol{\Sigma}|}{\sqrt{|\boldsymbol{\Sigma_1}||\boldsymbol{\Sigma_2}|}}\right)$ $\qquad \boldsymbol{\Sigma} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$

Color histograms

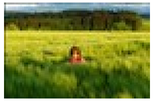
- General non-parametric model
  - Gaussian distribution is single-modal
  - Images are usually multi-modal



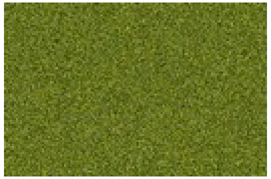


Histogram properties

- Robustness
  - Scale change, rotation
  - Resolution change
  - Partial occlusions



- No spatial information



- Sensitivity to illumination variation
  - Remove the value part

What is a texture?

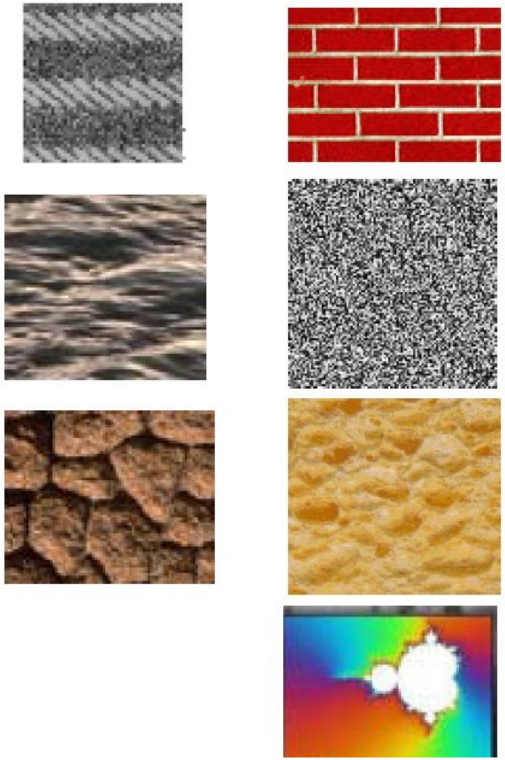
- No exact definition

  »Texture is a description of the spatial arrangement of color or intensities in an image or a selected region of an image.«

- Shape and texture
- Level of detail

Querying using texture

- Low-level description
  - Spatial properties
  - Frequency properties
- Perceptual properties
  - periodicity, coarseness, dominant orientation, complexity



repeatability

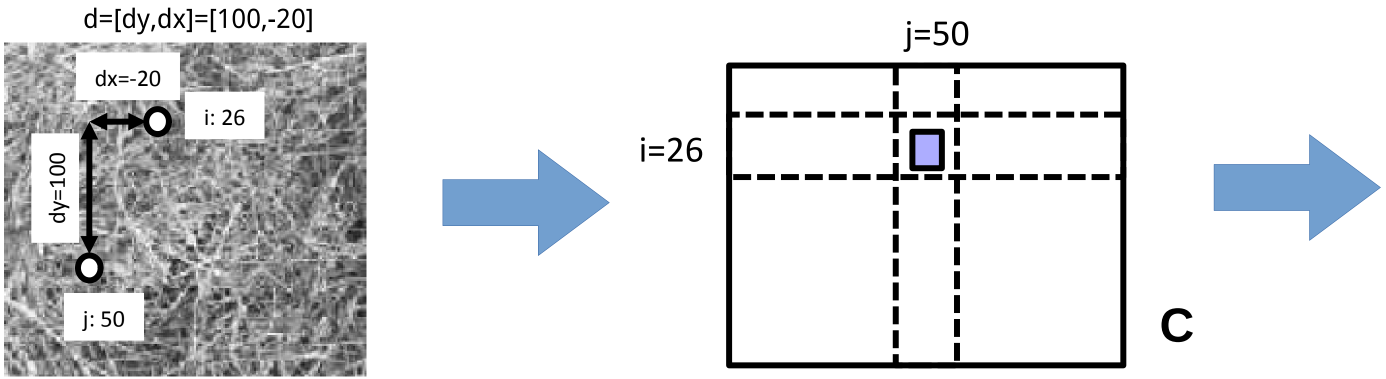
stochasticity

combination

fractals

Coocurrence matrix

- How many times does pixel of value V1 appear next to pixel of value V2?
  - Displacement vector d=[dy,dx]
  - C(i,j) contains number of times values i an j appear on image in relation d
  - Cooccurence matrix is normalized



v(**C**) = [f1,f2,f3,f4,f5,...]

Extracting features

Various features can be computed from cooccurence matrix

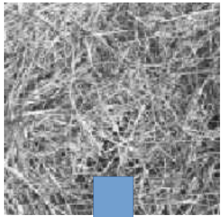
$$Energy = \sum_{i,j} C_A(i,j)^2$$

$$Entropy = -\sum_{i,j} C_A(i,j) log_2 C(i,j)$$

$$Contrast = \sum_{i,j} C_A(i,j)(i-j)^2$$

$$Homogenity = \sum_{i,j} \frac{C_A(i,j)}{1+|i-j|}$$

$$Correlation = \frac{\sum_{i,j}(i-\mu_i)(j-\mu_j)C_A(i,j)}{\sigma_i \sigma_j}$$



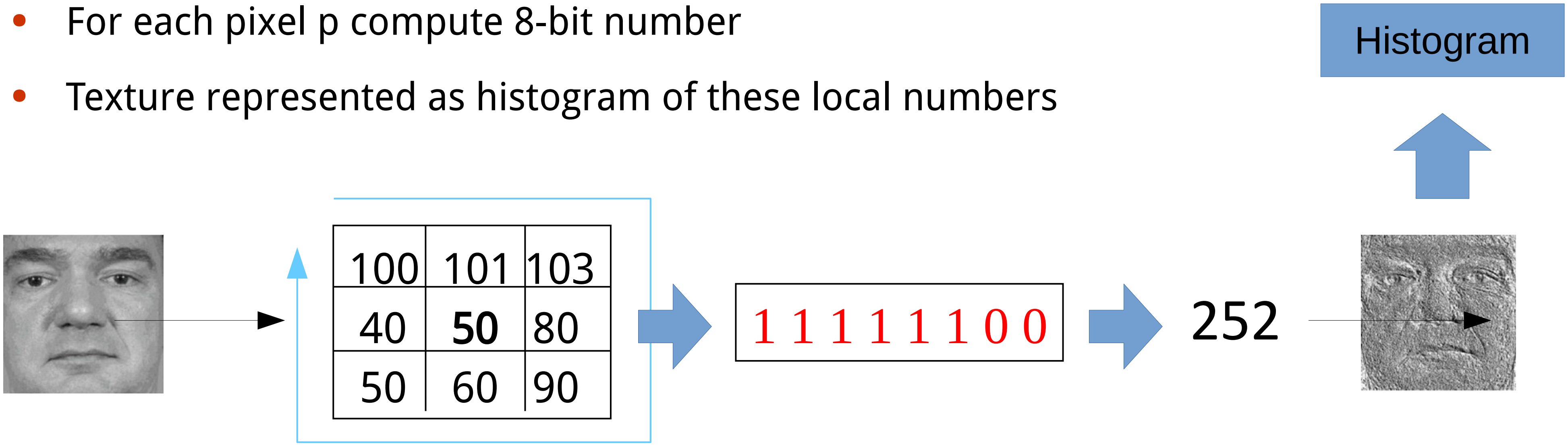
C

v(**C**) = [f1,f2,f3,f4,f5,...]

Comparison: Euclidean distance

Local Binary Pattern

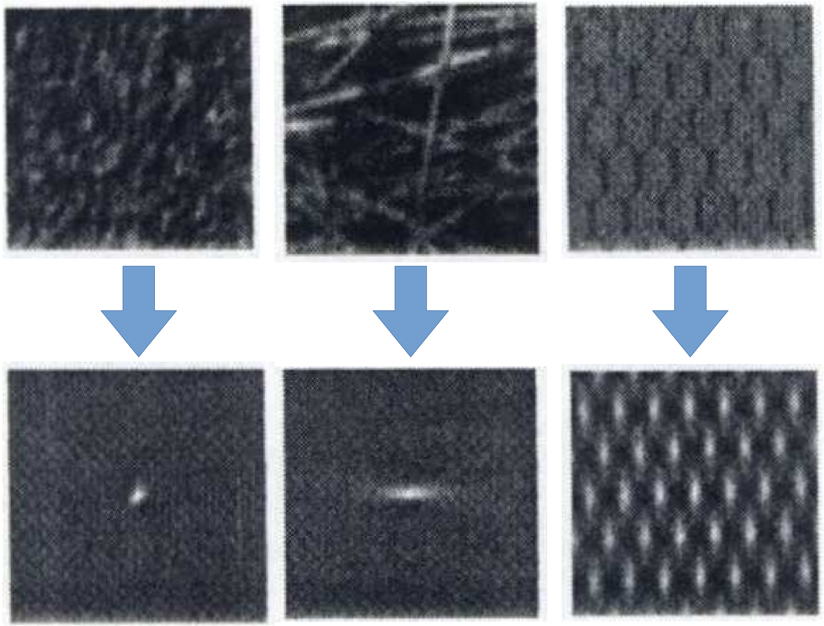
- Describe global texture with local descriptors
- For each pixel p compute 8-bit number
- Texture represented as histogram of these local numbers



T. Ahonen, A. Hadid, M. Pietikinen, "Face Description with Local Binary Patterns: Application to Face Recognition", TPAMI2006

Auto-correlation

- Normalized scalar product between image and its shifted version
- Shape of response function describes
  - Texture regularity
  - Texture coarseness



$$\rho(x, y) = \frac{\sum_{u,v} I(u, v) I(u + x, v + y)}{\sum_{u,v} I(u, v)^2}$$

Fourier transform

- Description of image with complex basis functions
  - Energy of spectrum: |F(u,v)|
  - If I is WxH, then F is WxH

$$F(u,v) = \mathcal{F}\{I(x,y)\}(u,v) = \frac{1}{WH}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} I(x,y)e^{-i2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)}$$

(0,0) *I(x,y)*

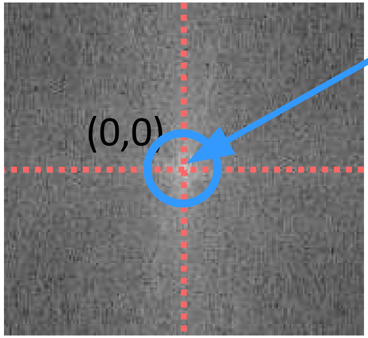
(W,H)

F(u,v) = 2dFFT[I(x,y)]

log(abs(F(u,v))+1)

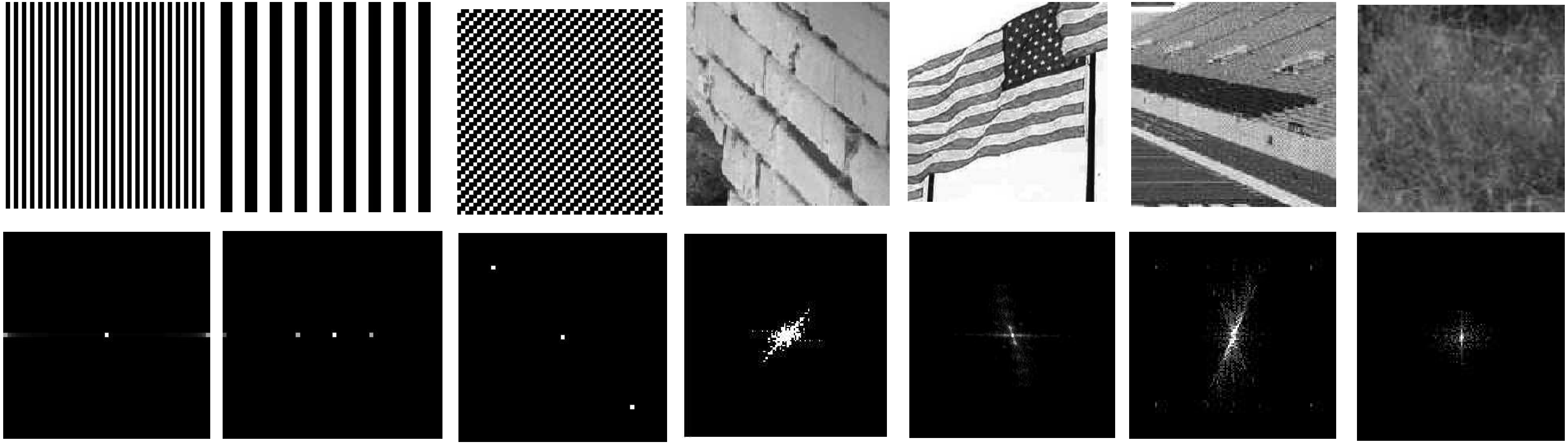
F(u,v)

Low frequencies



(u_{max},v_{max})

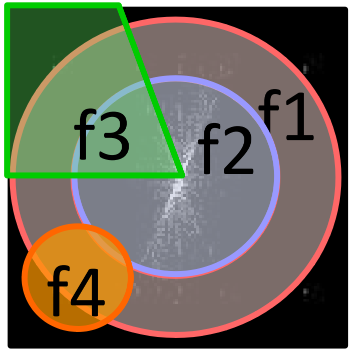
Examples



Spectrum features

How much energy is contained in various parts of spectrum

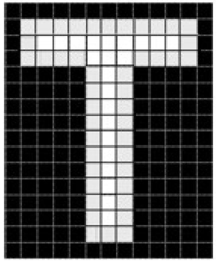
2D FFT

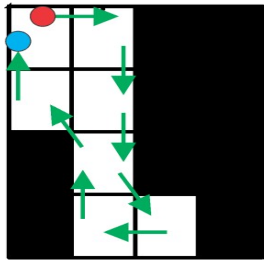


v = [f1, f2,f3,f4]

Query by shape

- Edge detection, threshold
- Vector of features
  - Region moments
  - Freeman differential codes
- Transformation distance
  - Amount of transformation



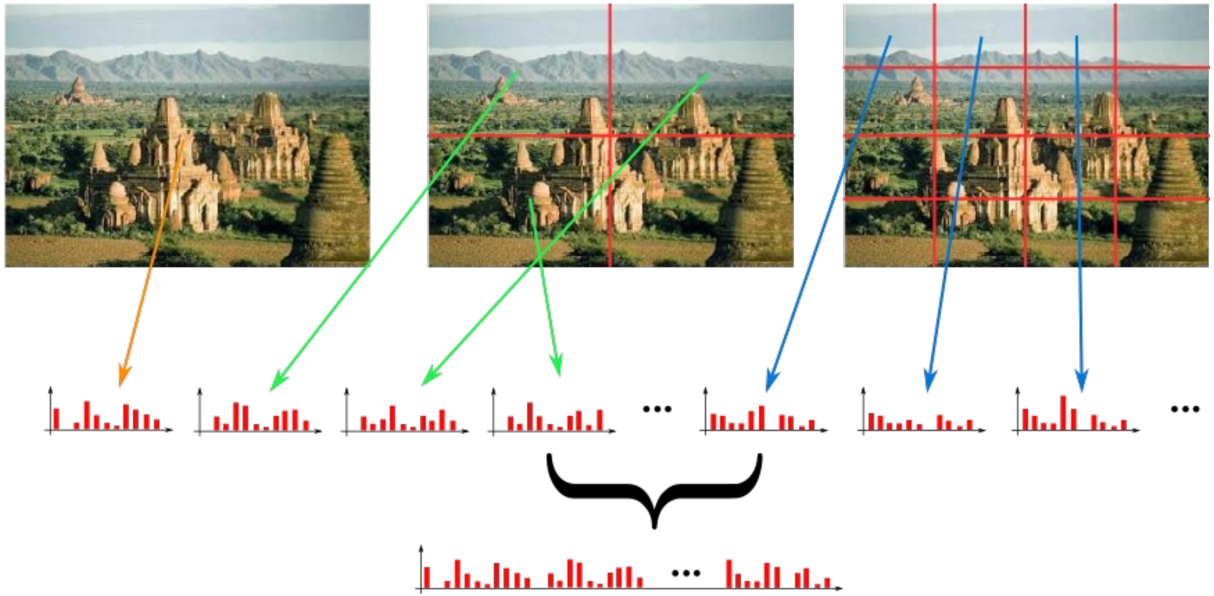


Comparing histograms

- Euclidean distance $D = \sqrt{\sum (h_1(i) - h_2(i))^2}$
- Hellinger distance $H = (\frac{1}{2} \sum_{i=1:N_{bins}} (h_1(i)^{\frac{1}{2}} - h_2(i)^{\frac{1}{2}})^2)^{\frac{1}{2}}$
- Chi-square distance $\chi^2 = \frac{1}{2} \sum_{i=1:N_{bins}} \frac{(h_1(i) - h_2(i))^2}{h1(i) + h2(i) + \varepsilon_0}$
- Histogram intersection $I = 1 - \sum_{i=1:N_{bins}} \min(h_1(i), h_2(i))$

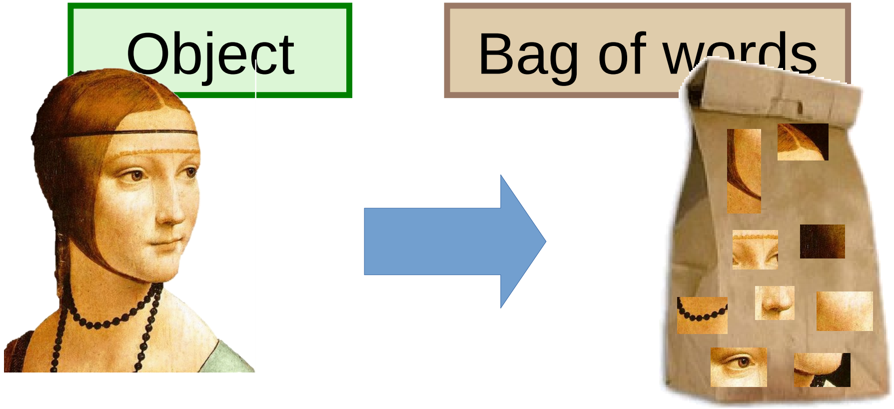
Including spatial information

- Divide image into subregions
- Stack histograms



Bag of words

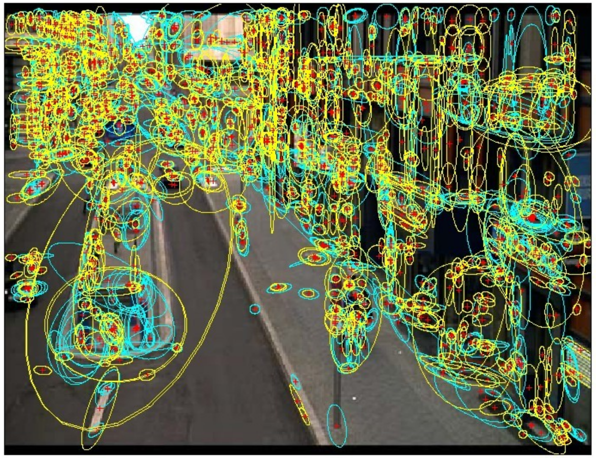
- Inspired by text retrieval systems
- General object categories
  - No clear spatial consistency
  - Objects composed of important parts - words
- Ignoring relationships between parts
  - Dictionary – list of known parts
  - Descriptor – histogram of part occurrences



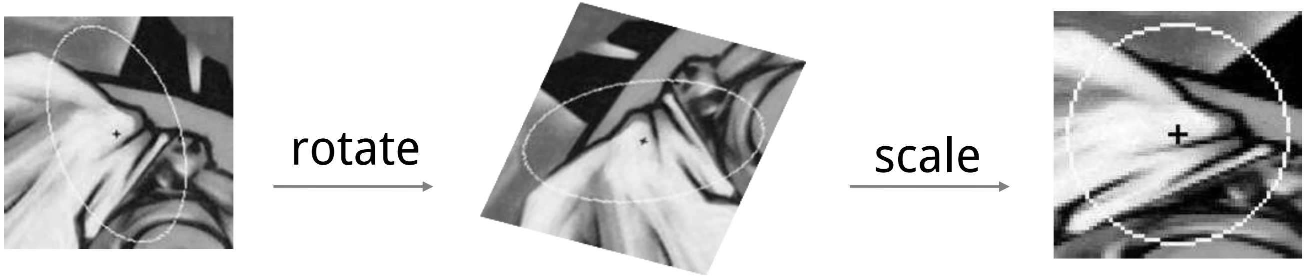
Visual words

Word	Feature
Token	Centroid/Cluster
Document	Image/Frame
Corpus	Video/Collection

Local regions

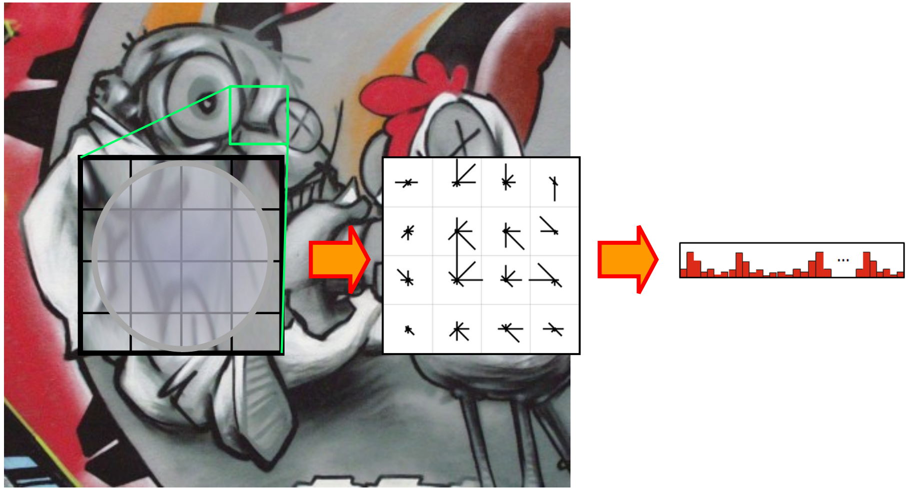


- Detecting stable regions
  - Robustness
  - Corners, blobs
- Describing neighborhood
  - Invariance (illumination, rotation, scale)



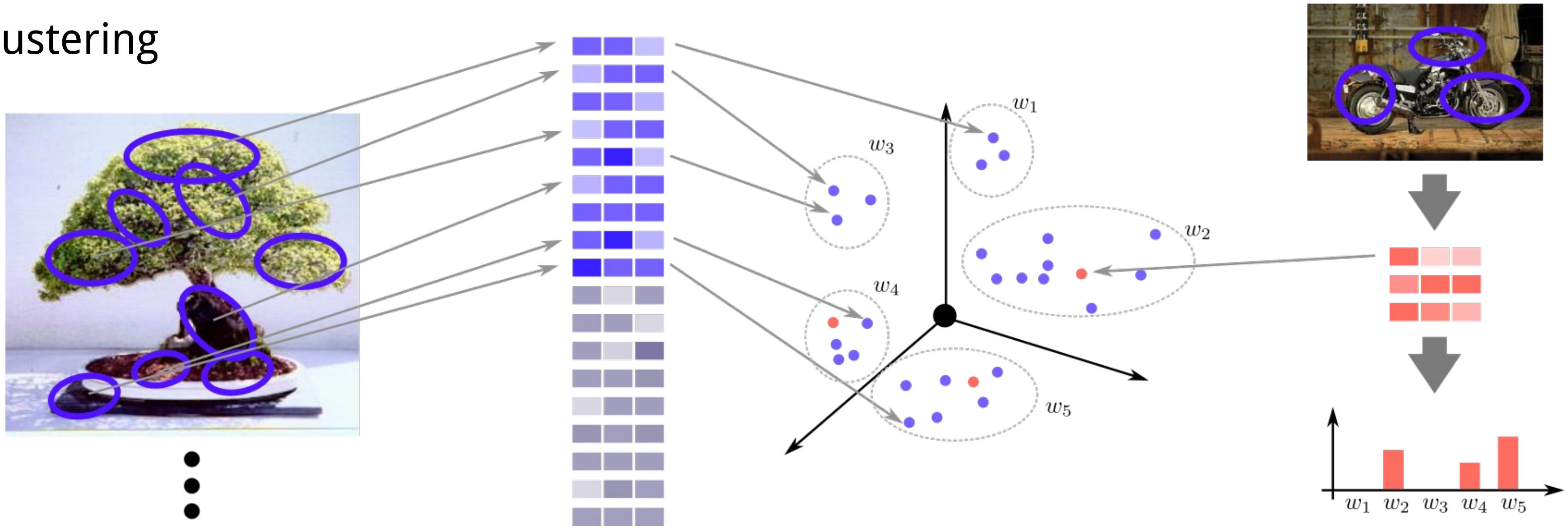
SIFT features

- Scale invariant feature transform
  - Divide region into 4x4 sub-regions: 16 cells
  - Compute gradients in each sub-region
  - Discretize orientation (8 directions)
  - Compute orientation histogram based on magnitude
  - Stack histograms and normalize: 4x4x8 = 128

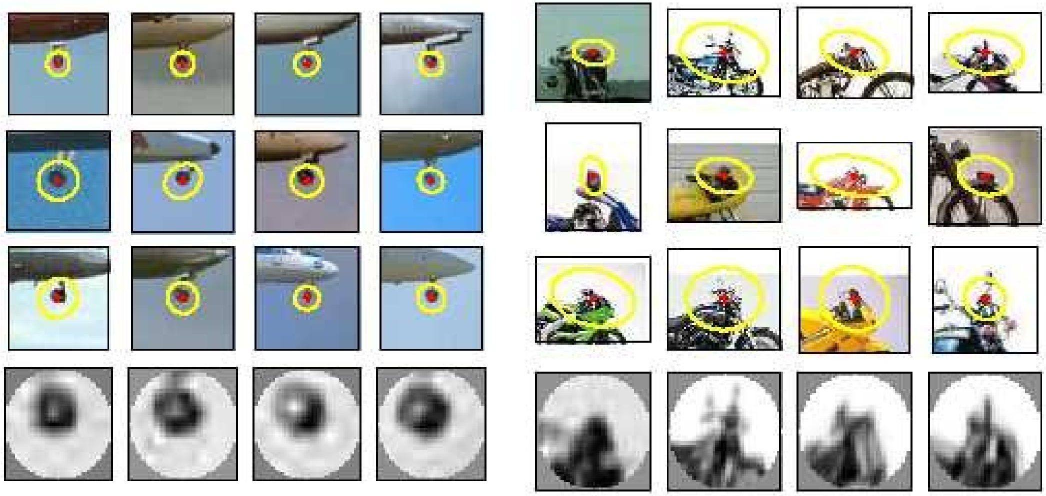


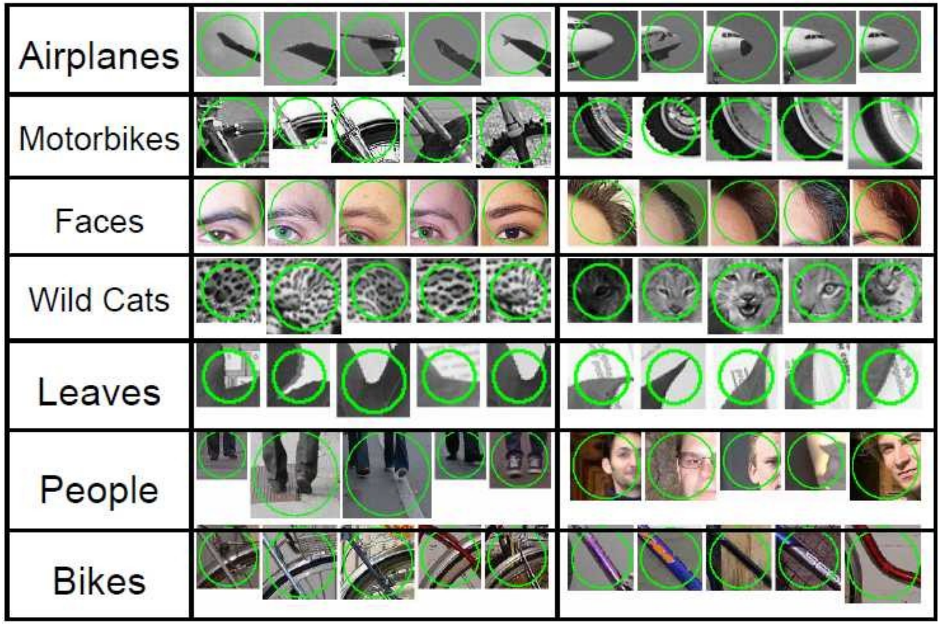
Building a dictionary

- Unsupervised learning
  - Large number of different local descriptors
  - Finite amount of words
  - Clustering



Example of visual words

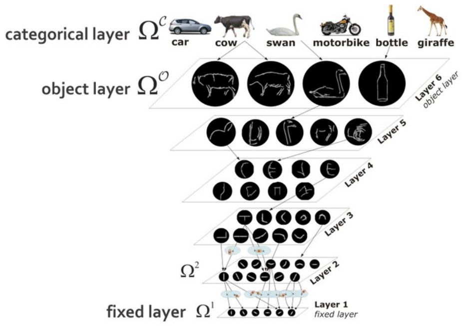




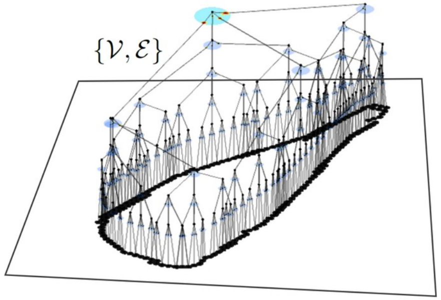
Sivic, Josef, and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos." IEEE CVPR, 2003

Hierarchy of parts

- Learn complex shape features
  - Gabor features – edges
  - Cooccurence
- Hierarchical composition
- Histogram of parts



Hierarchically organized vocabulary

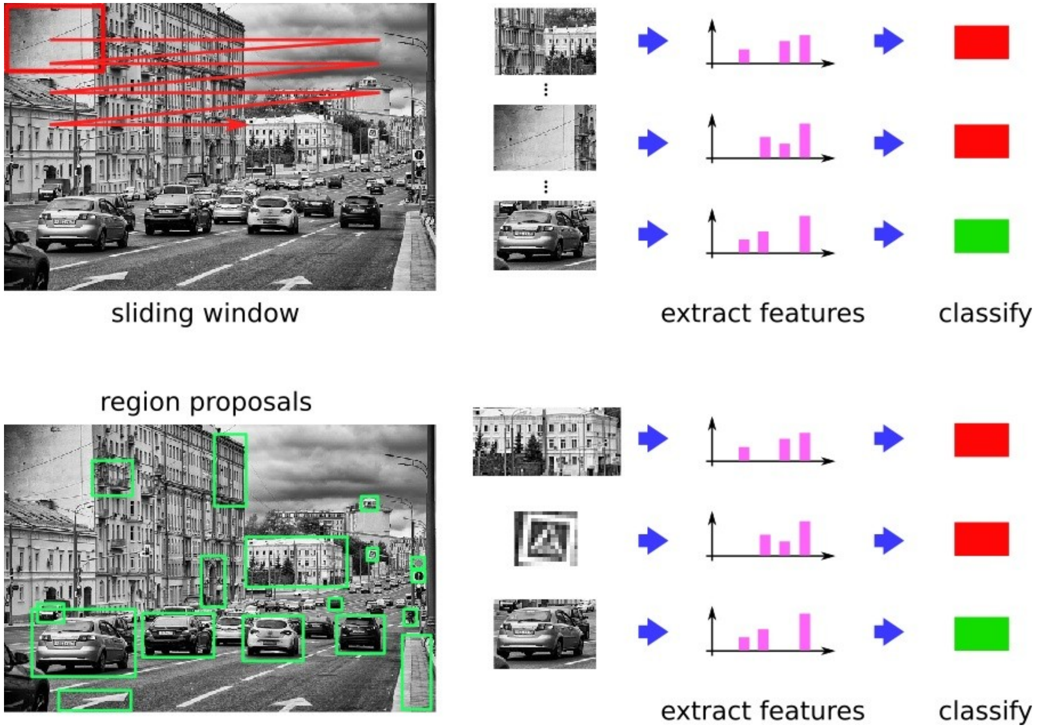


Example of a parse tree at detection

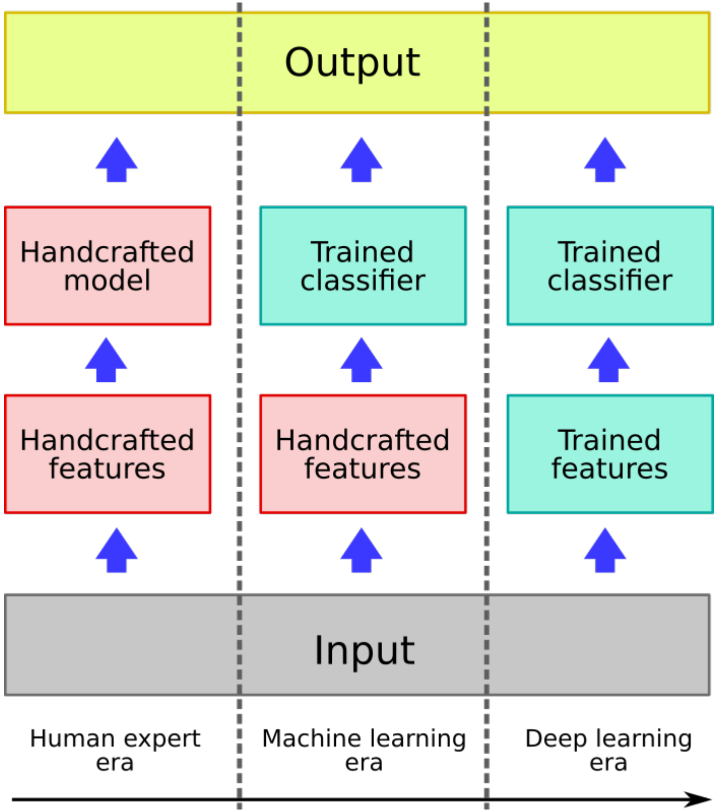
Sanja Fidler and Aleš Leonardis: Towards Scalable Representations of Visual Categories: Learning a Hierarchy of parts, CVPR 2007

Towards high-level categories

- Objects in images
- Scanning image
  - Sliding window
  - Region proposals
- Categorization
  - Features + SVM
  - CNN



Deep learning



CNN example – VGG16

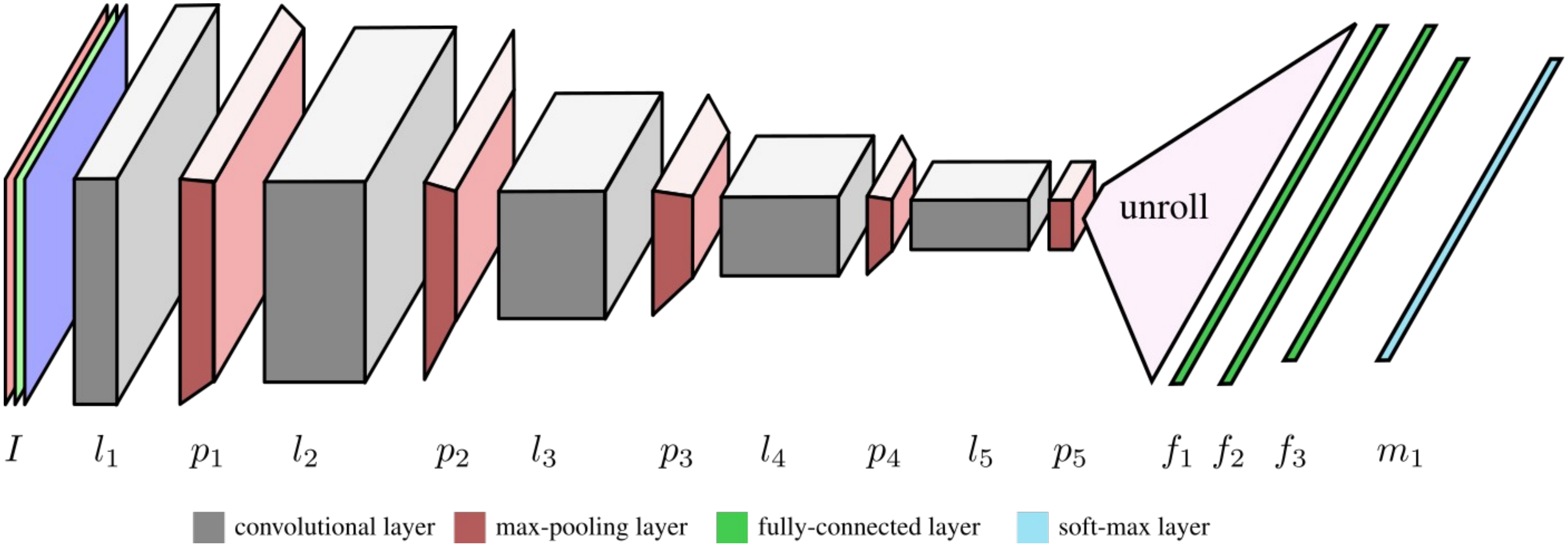
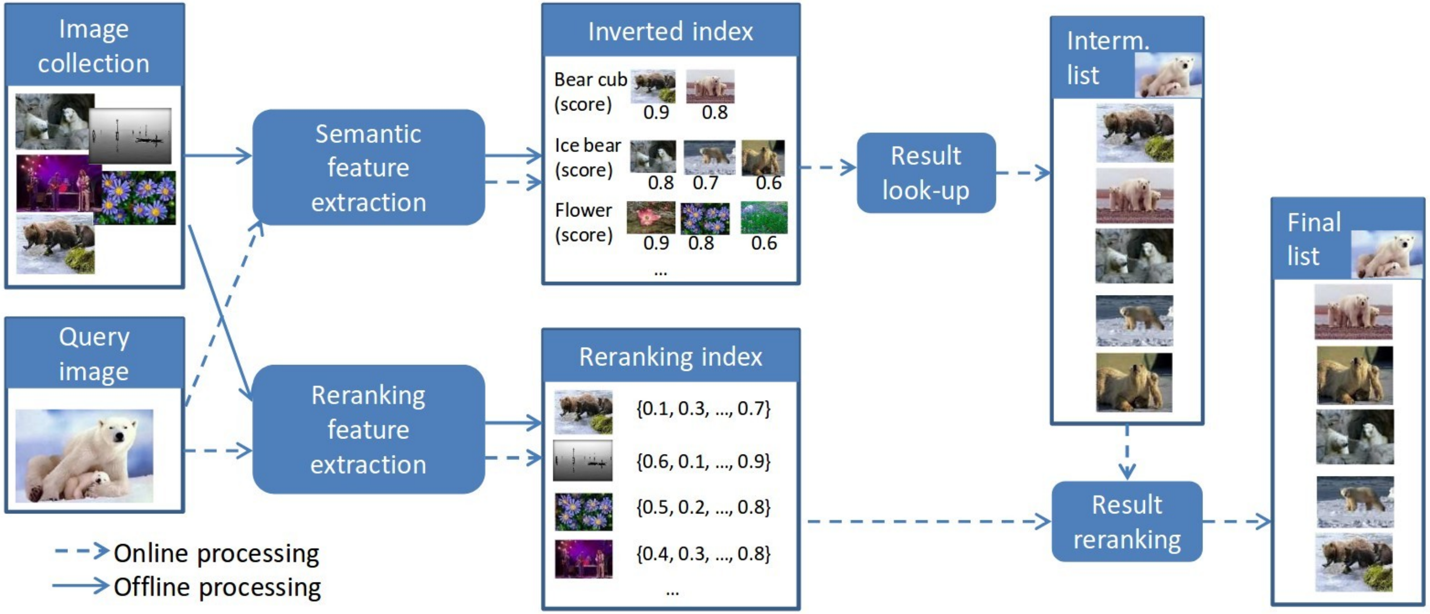


Image retrieval with inverted index

- Multi-object detector (semantic tokens)
- Use Boolean queries to per-process database



A. Popescu, A, Ginsca, H. Le Borgne, "Scale-free content based image retrieval (or nearly so)", ICCV 2017 Workshops

Efficient retrieval of dense descriptors

- Most descriptors are dense
  - Inverted index not efficient
  - Comparison is slow
- Structure the space
  - Hierarchical clustering
  - Traverse a tree (log n)
- Locality-sensitive hashing
  - Similar descriptors have the same hash value

Towards image understanding

- Semantic segmentation
- Spatial relationships
- Describing scene

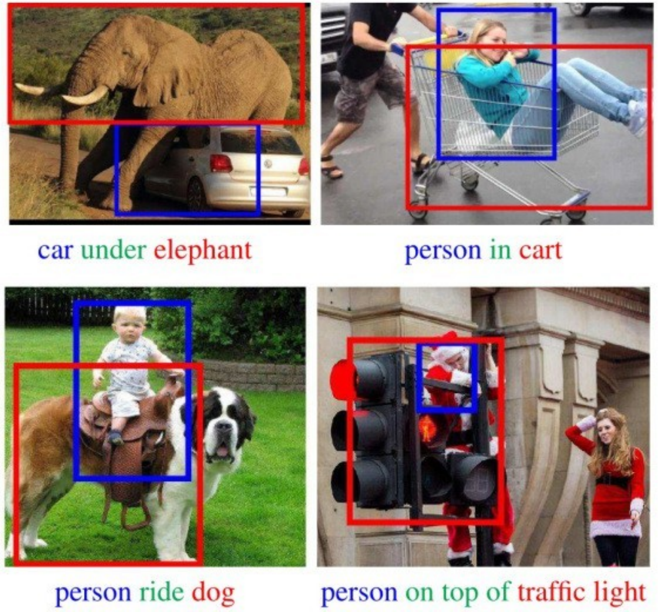
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

cs.stanford.edu/people/karpathy/deepimagesent/



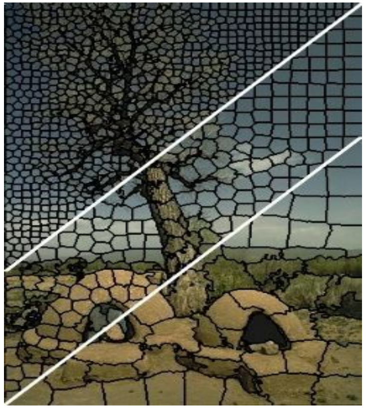
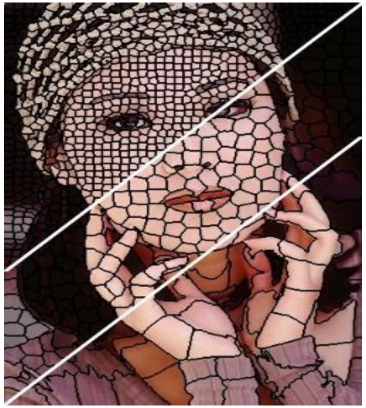
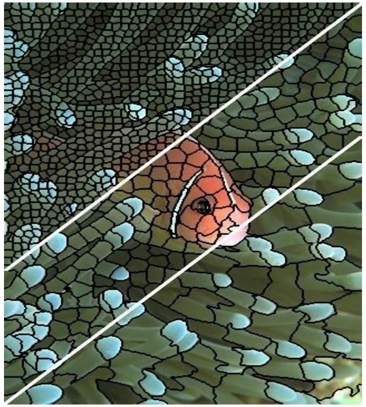
www.di.ens.fr/willow/research/unrel/

Why decompose images?

- Retrieval with specific queries (e.g. horses)
- Describe entire image
  - Which descriptors belong to object?
- Describe only parts of images
  - How many, what shape?

Superpixels

- Over-segmentation
- Describe each pixel in CIE Lab and (x,y): $C_i = [l_i \; a_i \; b_i \; x_i \; y_i]^{\mathrm{T}}$
- Manually set number of clusters (superpixes)
- Modified K-means (fast, spatial restrictions)

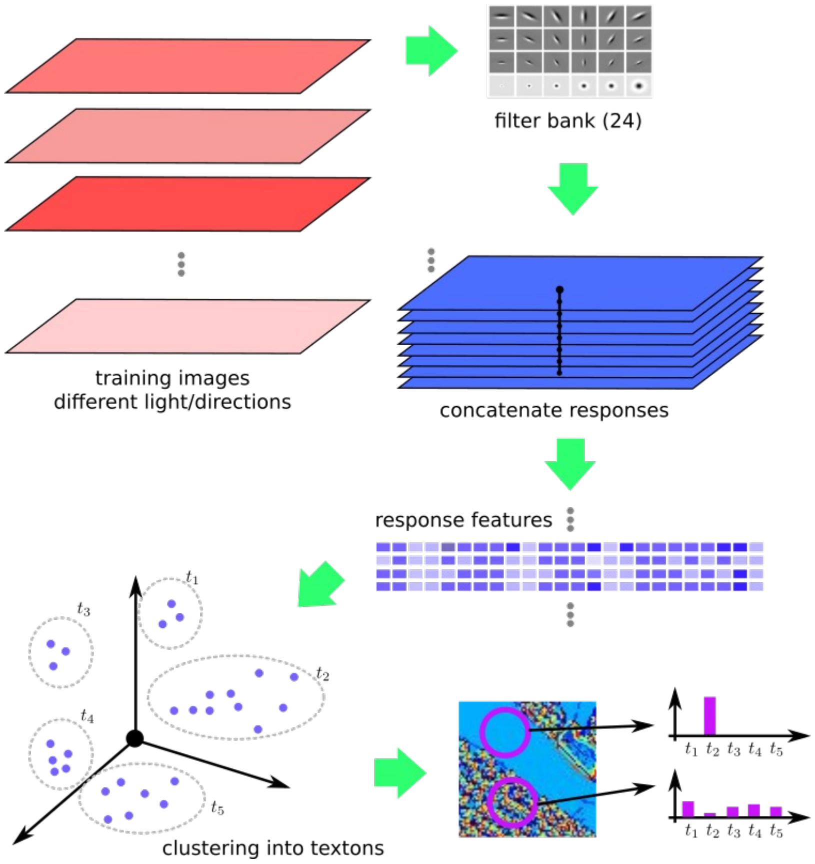
R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC Superpixels Compared to State-of-the-art Superpixel Methods, PAMI2012

Automatic decomposition examples

Li et al., Segmentation Using Superpixels: A Bipartite Graph Partitioning Approach, CVPR2012

Texton segmentation

- Texton descriptor learning
  - Each pixel described with responses to a bank of filters (e.g. 24 filters)
  - Find textons by clustering responses of filters
- Assign each pixel a texton
- Describe texture around pixel as a histogram of textons
- Segmentation - cluster histograms



T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV, 2001

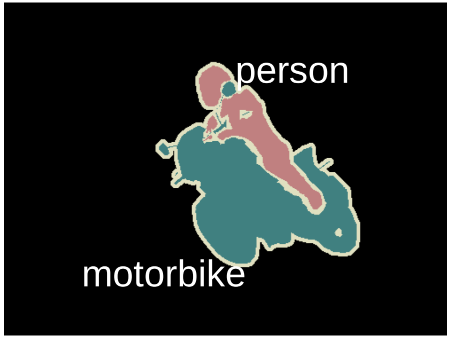
Segmentation using texture

Original image | k-means (k=5), feature: rgb | k-means (k=5), feature: texton

Multiclass segmentation using textons

Semantic segmentation

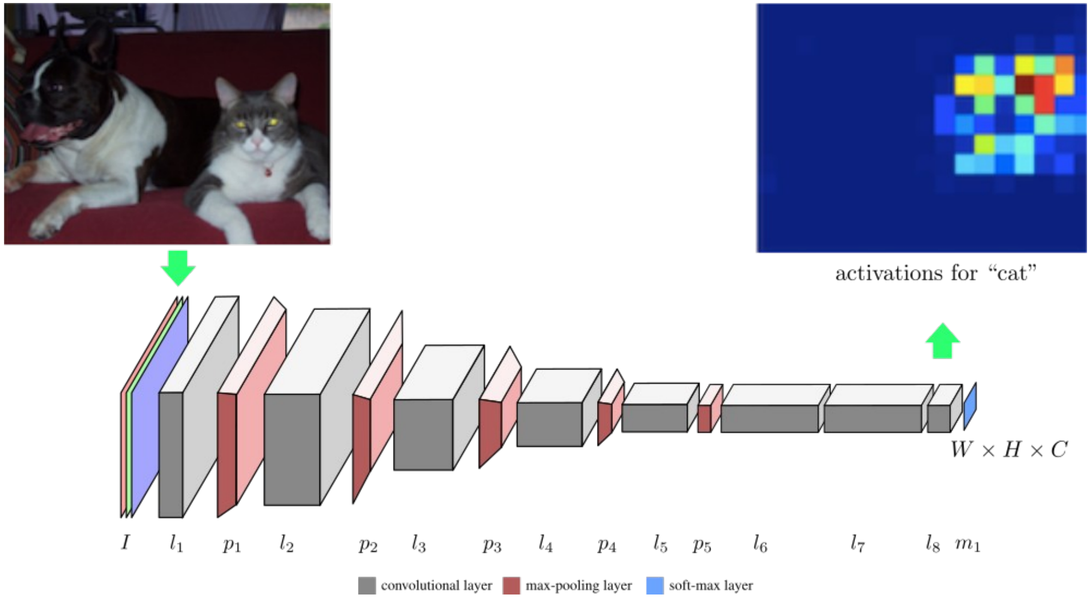
- Segments have semantic meaning
- "Bag-of-textons"
  - Texton features
  - Classifier



- Convolutional neural networks
  - Train network for per-pixel classification
  - Encoding context

Handling subsampling in CNNs

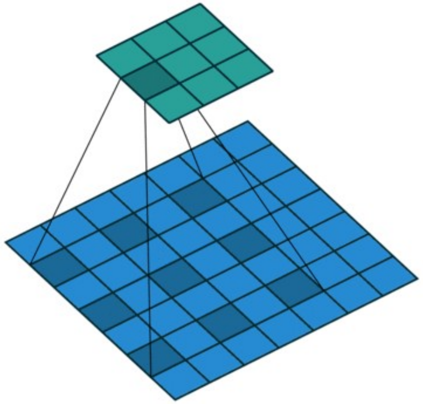
- Pooling/subsampling
  - Reduce parameter count
  - Increase spatial robustness
- Approaches
  - Interpolation
  - MRF
  - Deconvolution

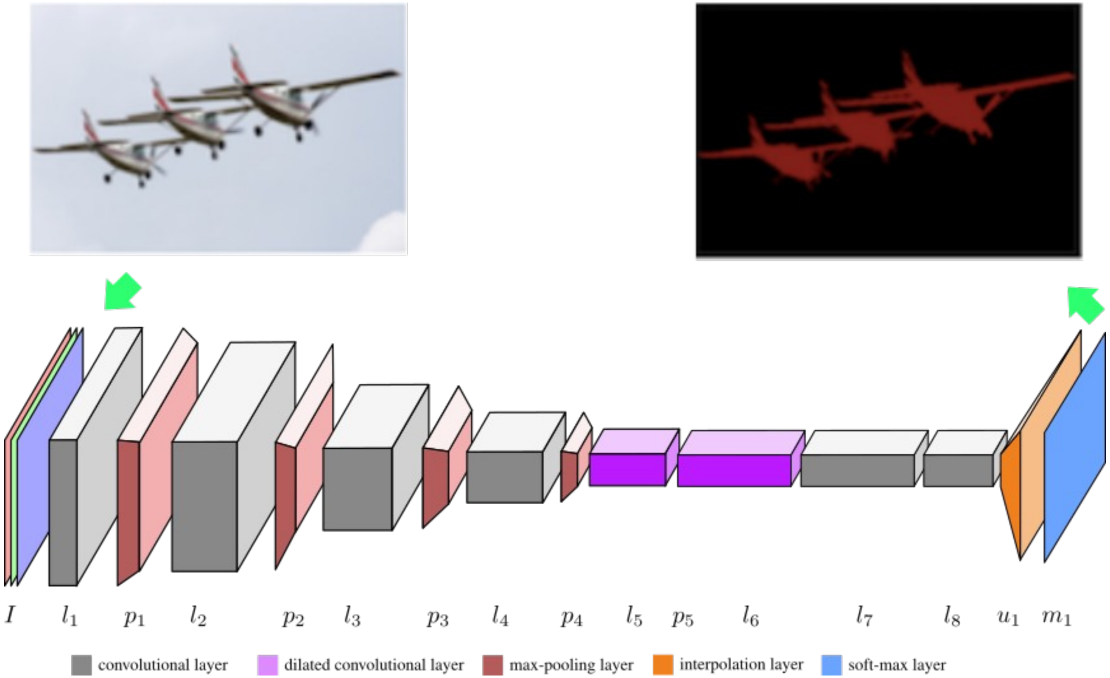


J, Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, IEEE TPAMI 2017

Avoiding pooling

- Dilated convolution
- MRF

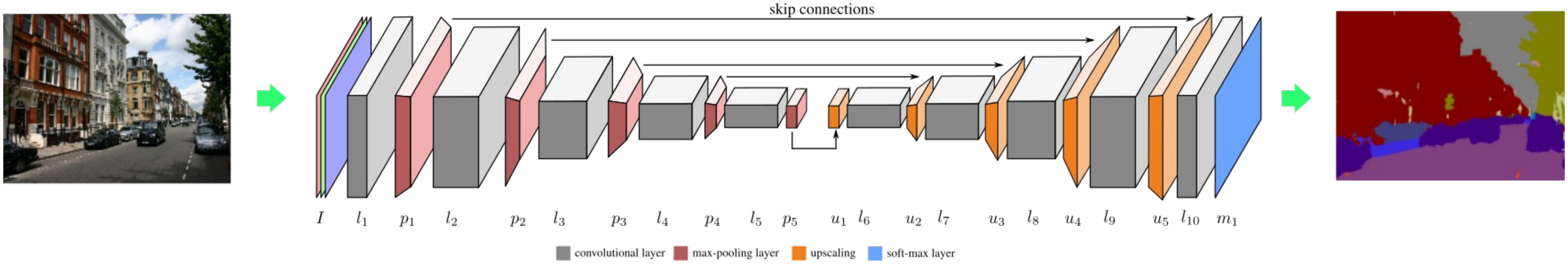




L.C. Chen et al., DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE TPAMI 2017

Encoder-decoder

- Deconvolution produces coarse segments
- Skip connections
  - Information from hi-res features



Describing video content

- Structure: frame, shot, scene
- Content
  - Dynamics: still, moving objects, camera movement
  - Activity in a frame interval, e.g. jumping, robbery, horse race
  - Categories, e.g. cats, horses, cars
  - Object instances: e.g. Harry Potter, Jack Sparrow, Han Solo

MPEG-7

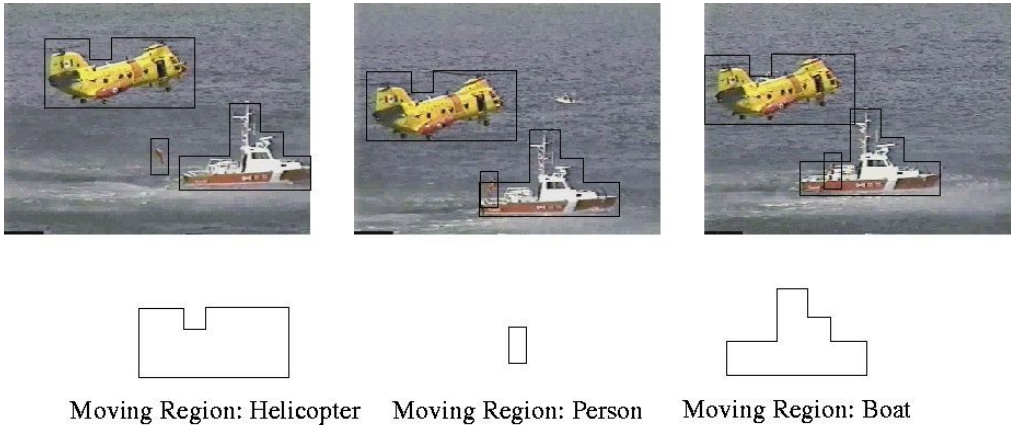
- Efficient access and manipulation of multimedia content
- Complementary to MPEG-4
- Standardized text-less object retrieval
  - D – Object descriptors (audio and video)
  - DS – Description schemes
  - DDL – Description definition language (XML)

Examples of descriptors

- Color
  - Color space
  - Color layout
  - Dominant color
  - Color structure
  - GoP color
- Texture
  - Homogenous
  - Non-homogenous
- Shape
  - Shape descriptor
  - Contour
  - 2D-3D shape
- Motion
  - Activity
  - Camera motion
  - Warping parameters
  - Trajectory
  - Parametric motion
- Localization
  - Spatio-temporal
  - Region

Structure description

Describing content at the level of video segment



Example: three moving objects, describe relations ...

Applications

- Digital library (Image/video/music catalogue)
- Broadcast media (Radio channel, TV channel)
- Multimedia authoring
- E-business: Searching for products
- Cultural services (art-galleries, museums)
- Educational applications
- Biomedical applications