Retrieval in multimedia

Overview

- Visual information retrieval
- Audio information retrieval
- Making retrieval efficient
  - Hierarchical methods
  - Vector databases

As text retrieval

- Documents can be queried using
  - Metadata (text)
  - User annotations (tags)
  - Manual annotations (tags, captions)
- Problems
  - Metadata is not complete/informative/available
  - User annotations not supported, unreliable
  - Captioning is selective / biased

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

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

Querying image content

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

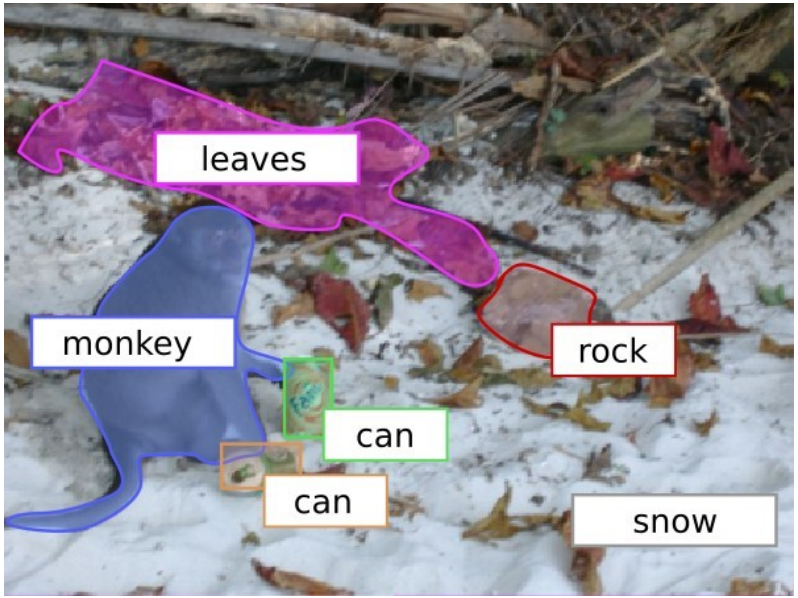
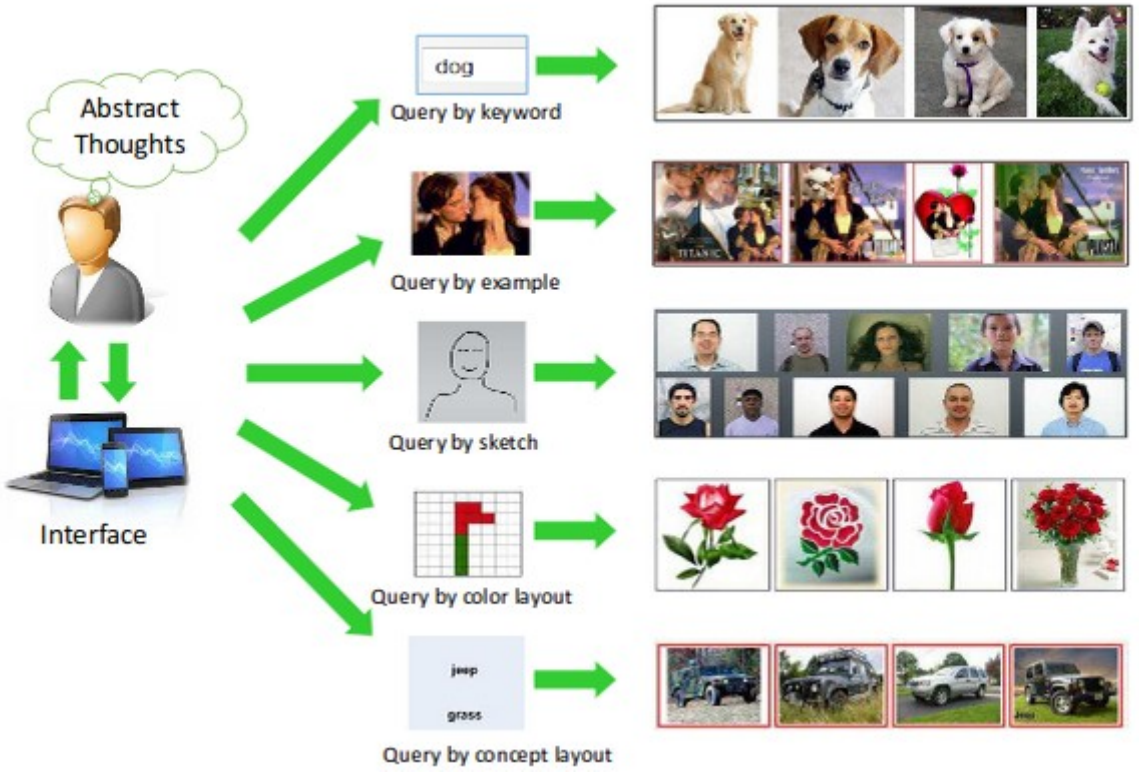


Image retrieval systems



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

Image retrieval system

Querying by color

- Average color
- Parametric distribution (Gaussian)
  - Signle mode
- Color histogram
  - Multi-modal
  - Illumination change sensitivity

μ_a μ_b

(μ_a, σ_a) (μ_b, σ_b)

$[a_1, a_2, \ldots]$ $[b_1, b_2, \ldots]$

What is a texture?

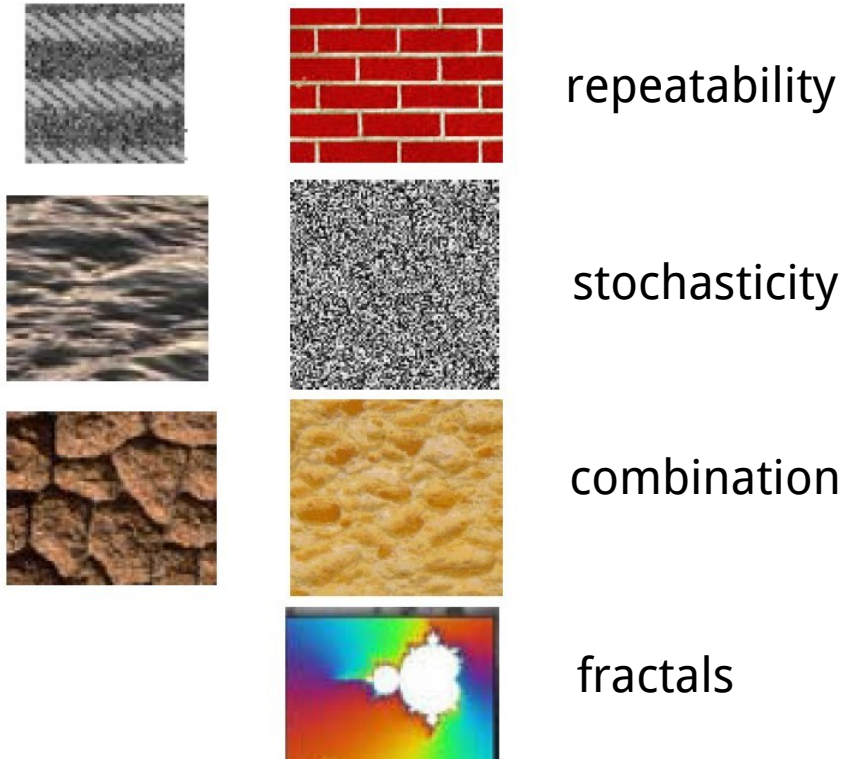
- No exact definition

  »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

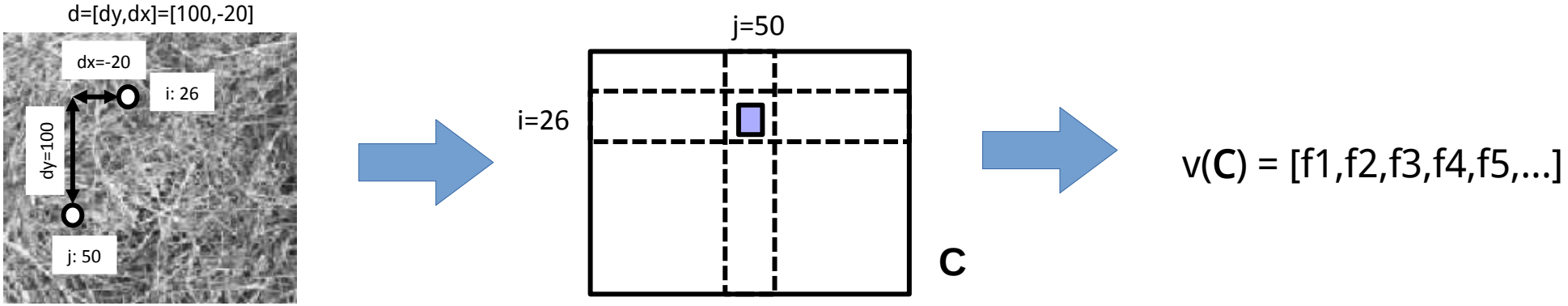
Querying using texture

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



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



Extracting features

Various features can be computed from cooccurence matrix

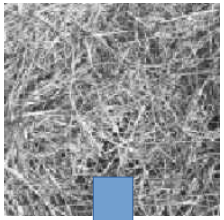
$$Energy = \sum_{i,j} C_A(i,j)^2$$

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

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

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

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



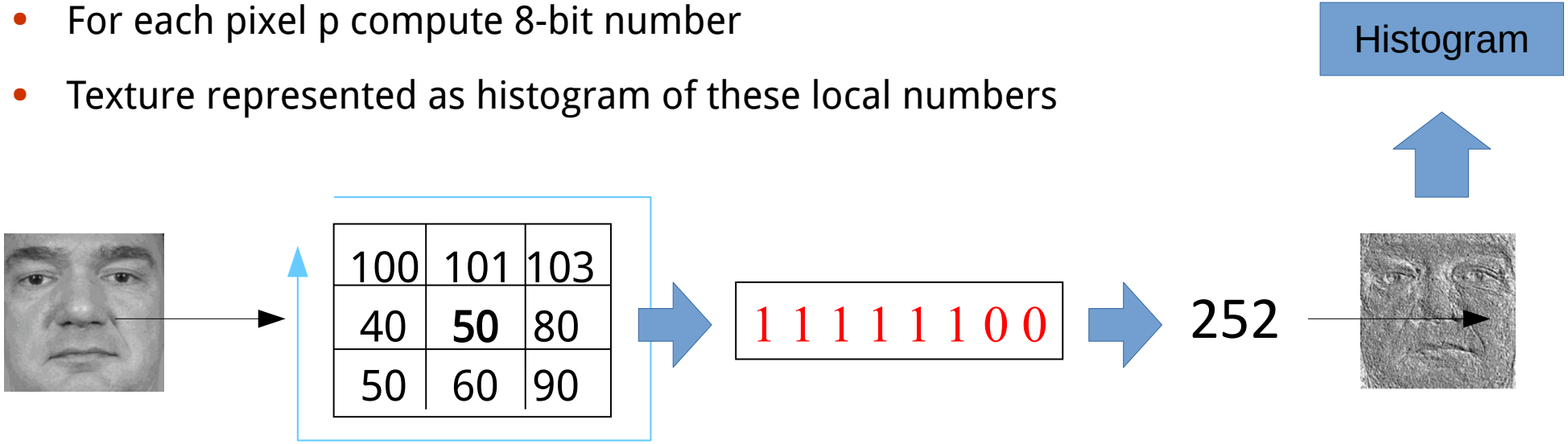
C

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

Comparison: Euclidean distance

Local Binary Pattern

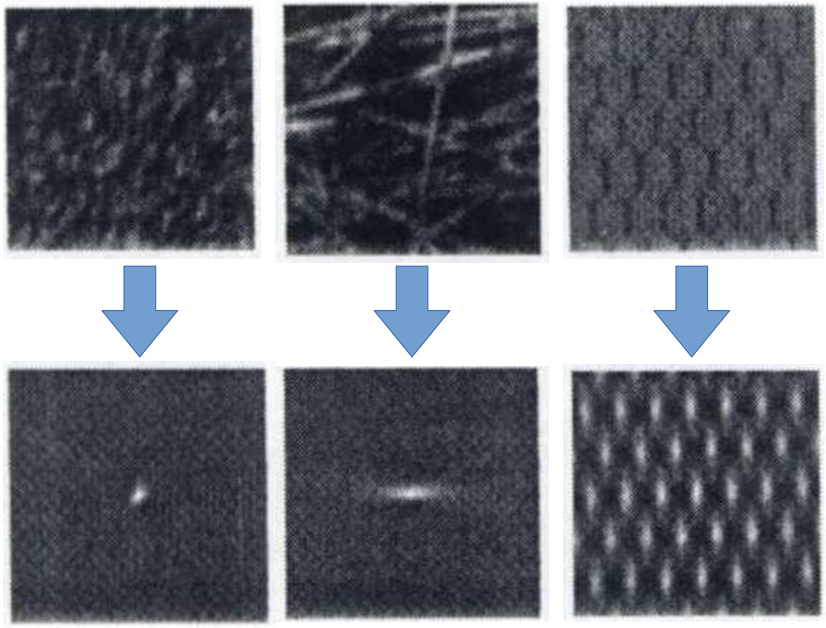
- Describe global texture with local descriptors
- For each pixel p compute 8-bit number
- Texture represented as histogram of these local numbers



Auto-correlation

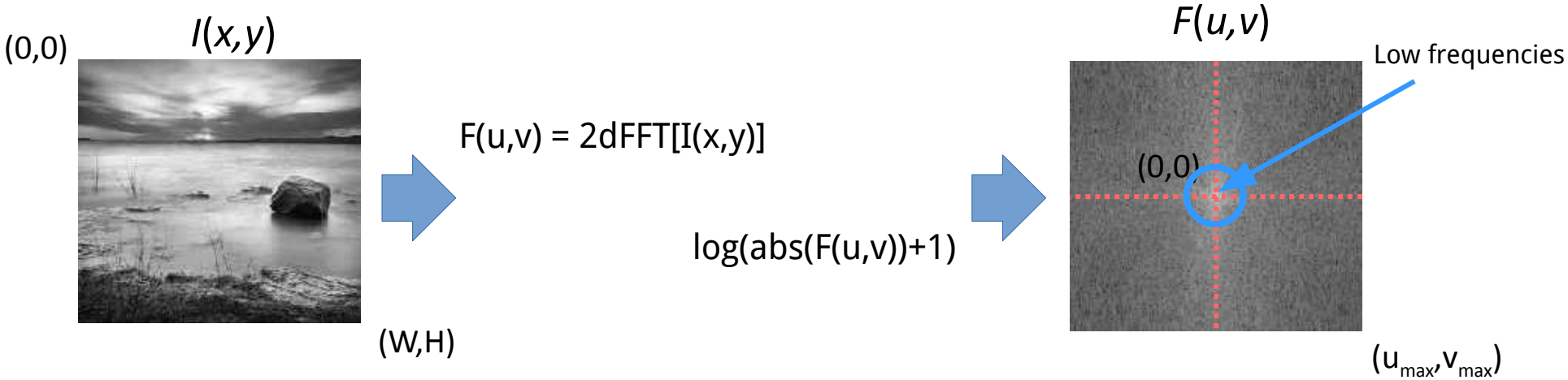
- Normalized scalar product between image and its shifted version
- Shape of response function describes
  - Texture regularity
  - Texture coarseness

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

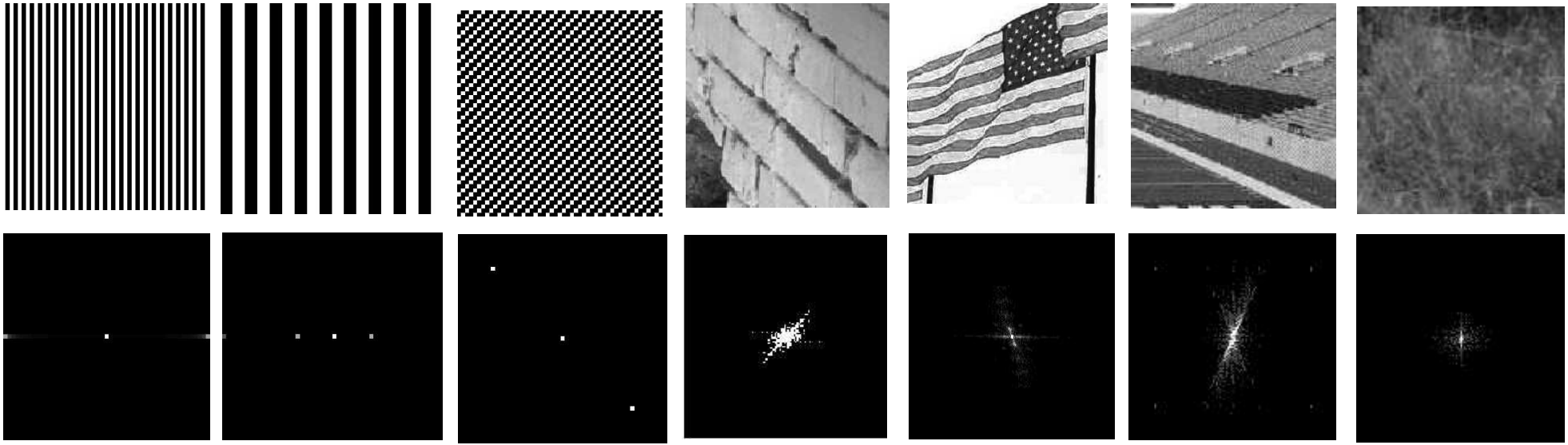


Fourier transform

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



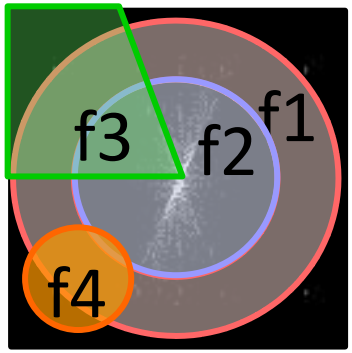
Examples



Spectrum features

How much energy is contained in various parts of spectrum

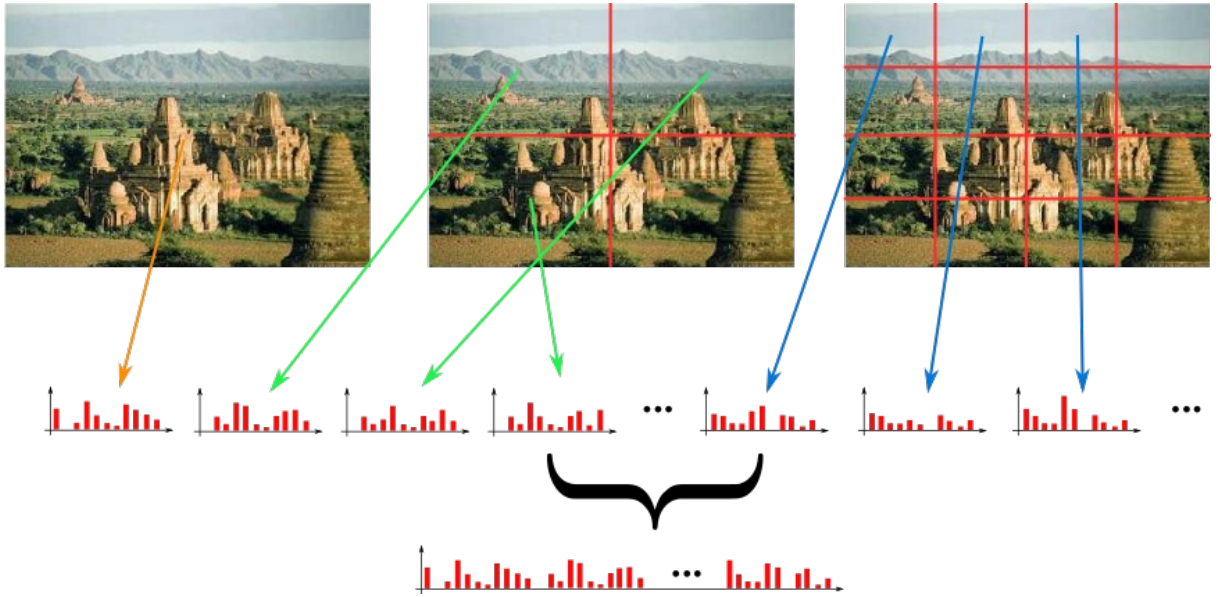
2D FFT



v = [f1, f2,f3,f4]

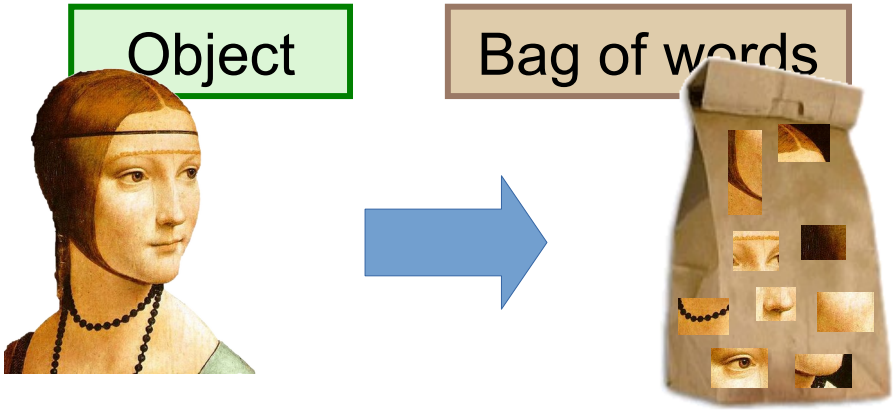
Including spatial information

- Divide image into sub-regions
- Stack histograms

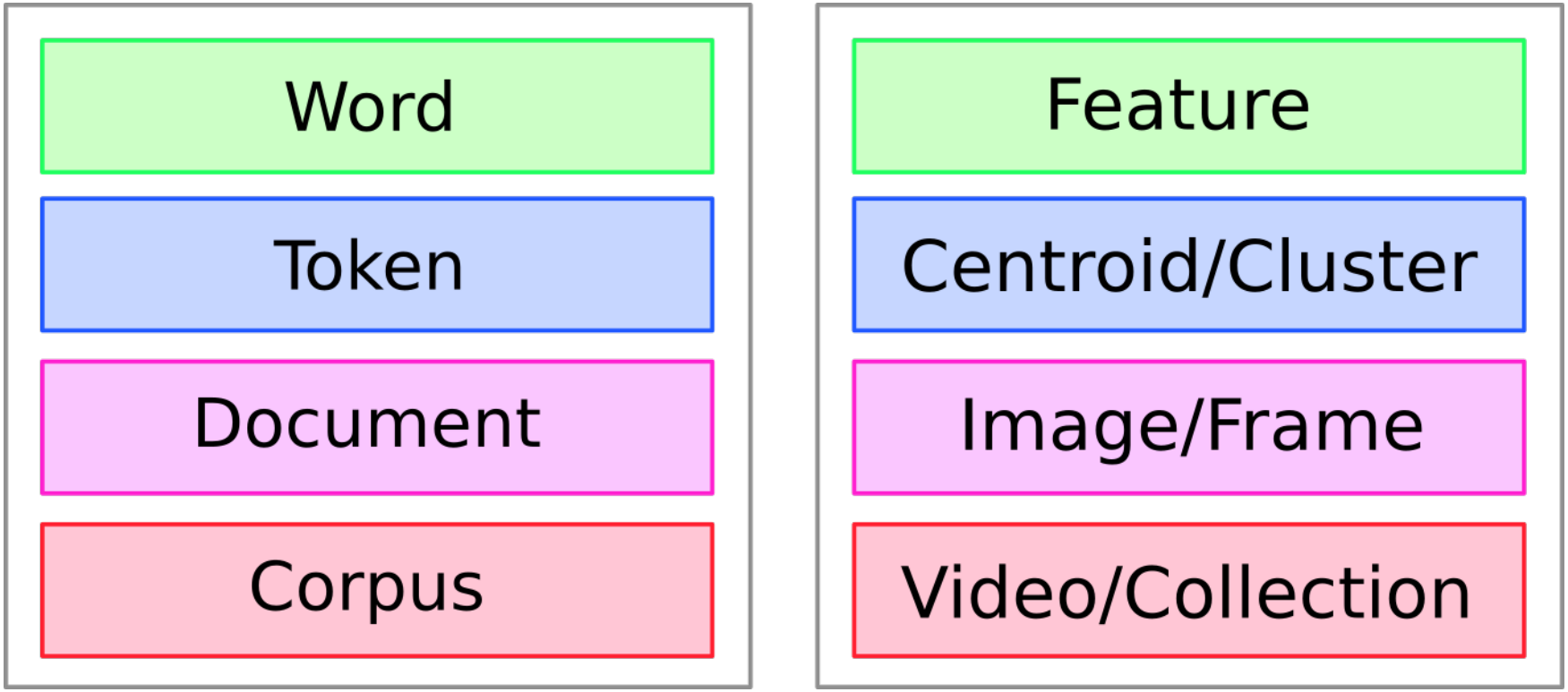


Bag of words

- 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

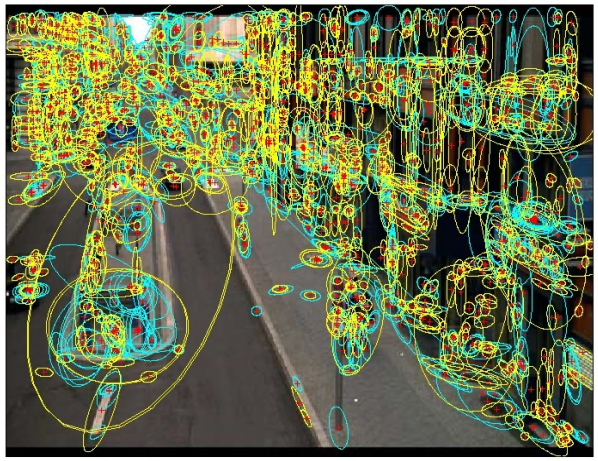


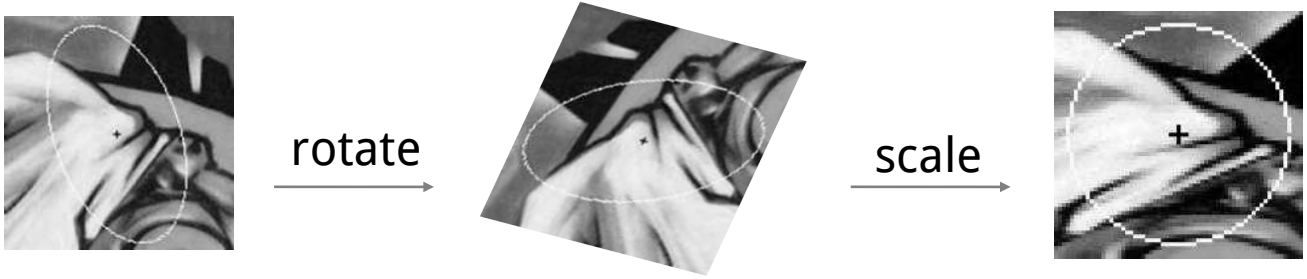
Visual words



Local regions

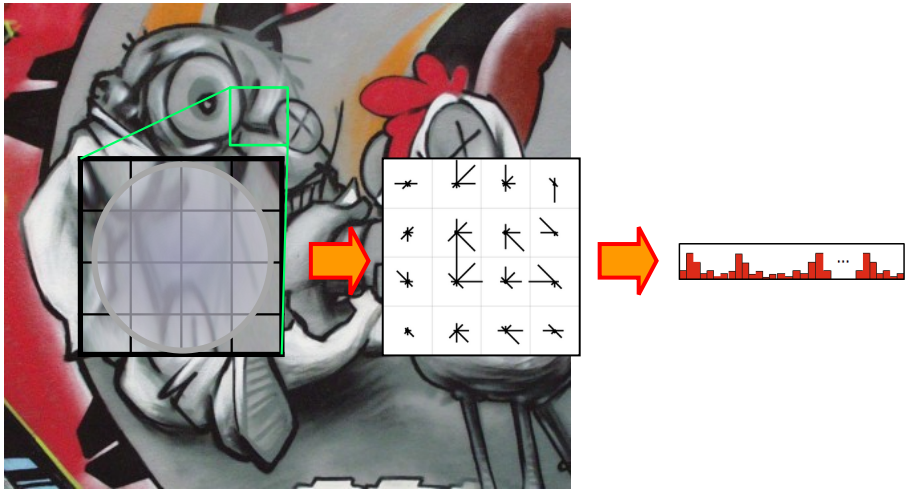
- Detecting stable regions
  - Robustness
  - Corners, blobs
- Describing neighborhood
  - Invariance (illumination, rotation, scale)





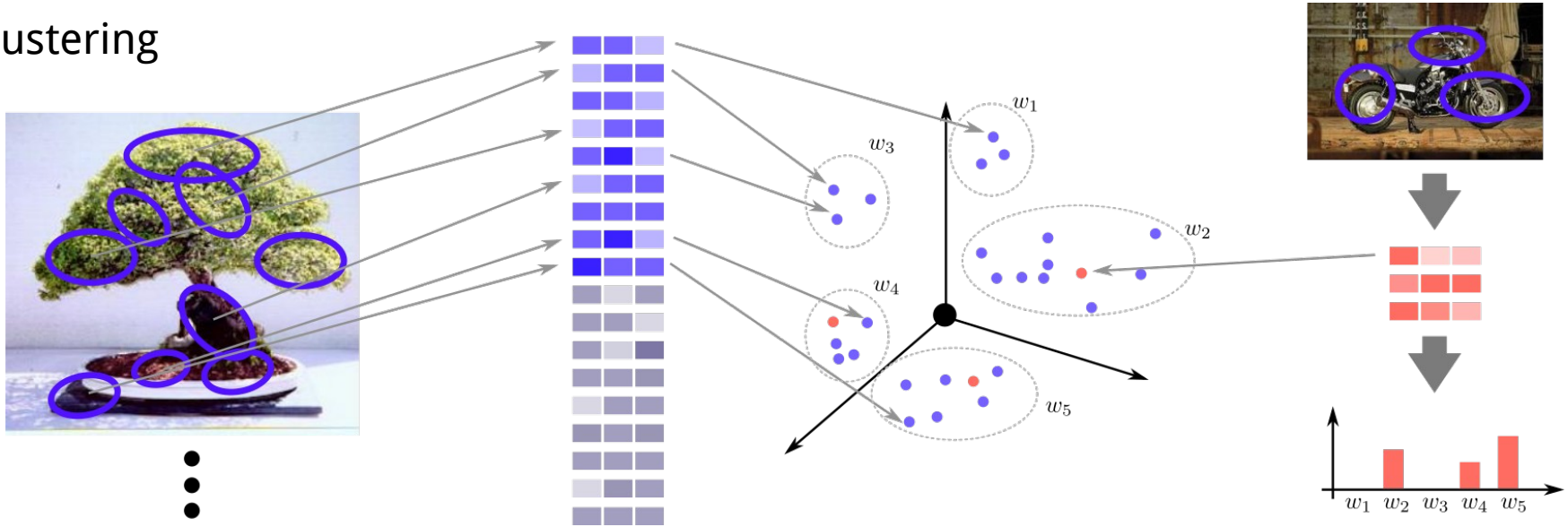
SIFT features

- 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

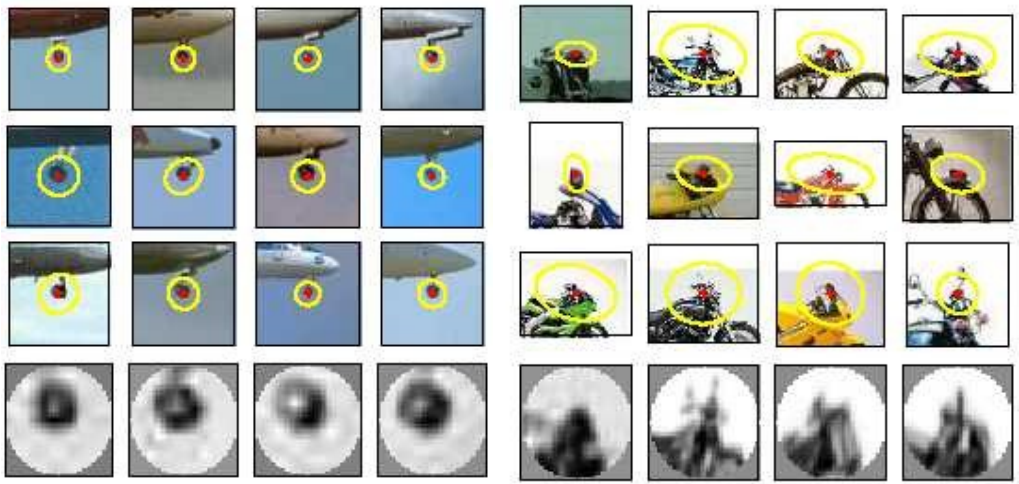


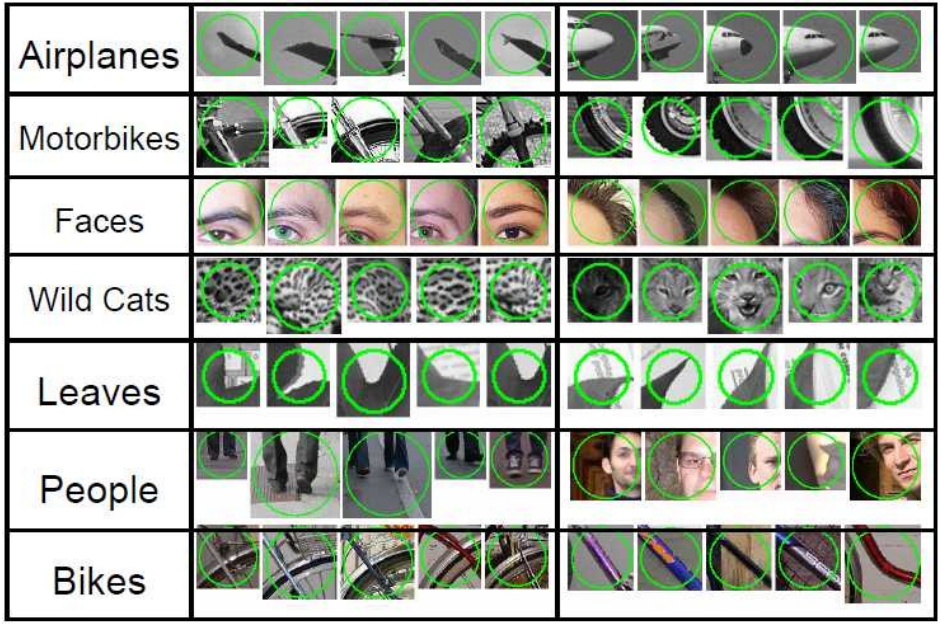
Building a dictionary

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



Example of visual words

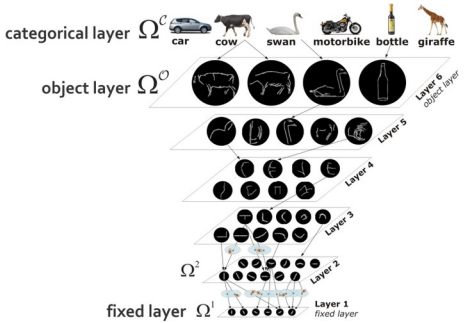




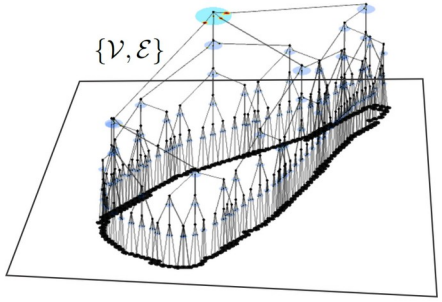
Sivic, Josef, and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos." IEEE CVPR, 2003

Hierarchy of parts

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

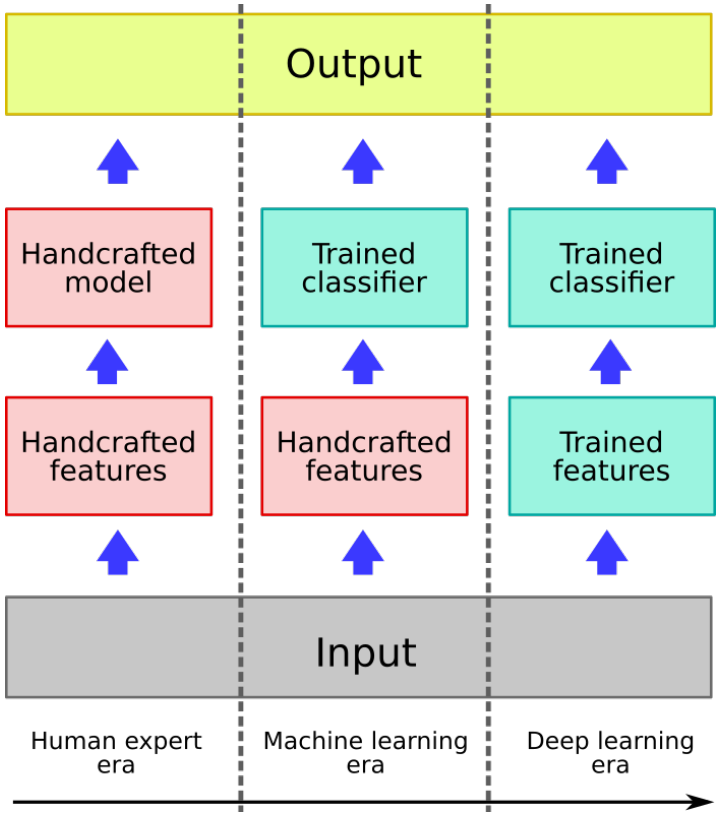


Hierarchically organized vocabulary

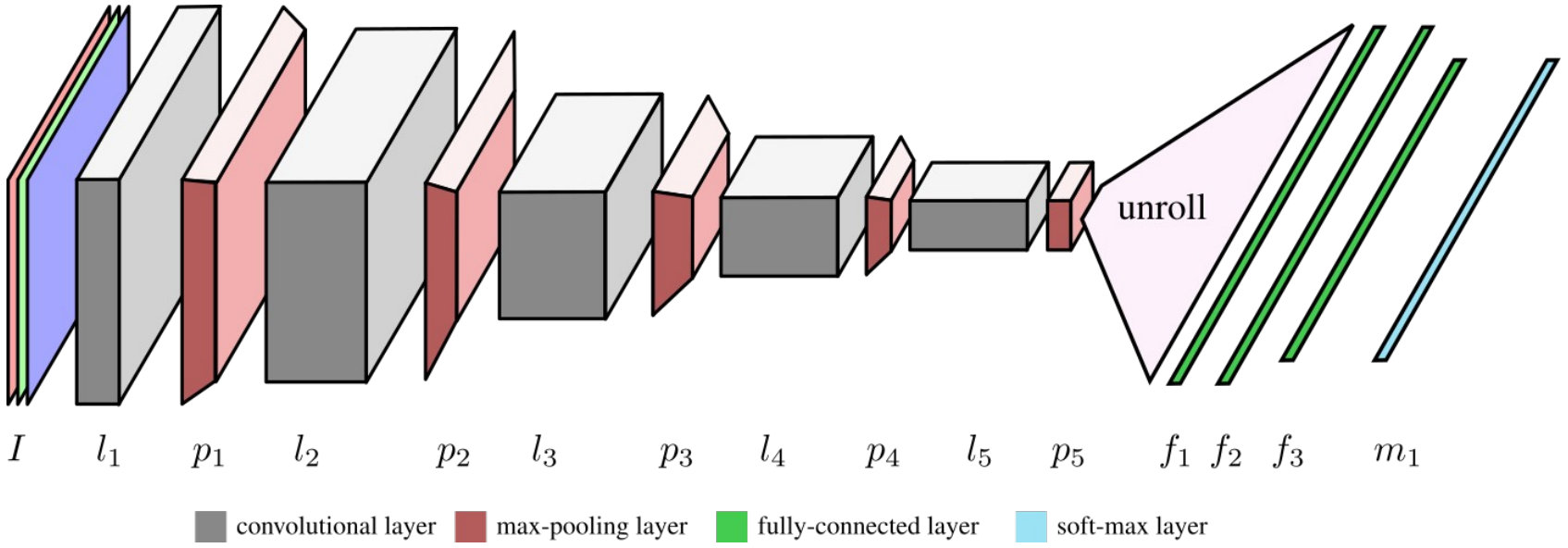


Example of a parse tree at detection

Deep learning



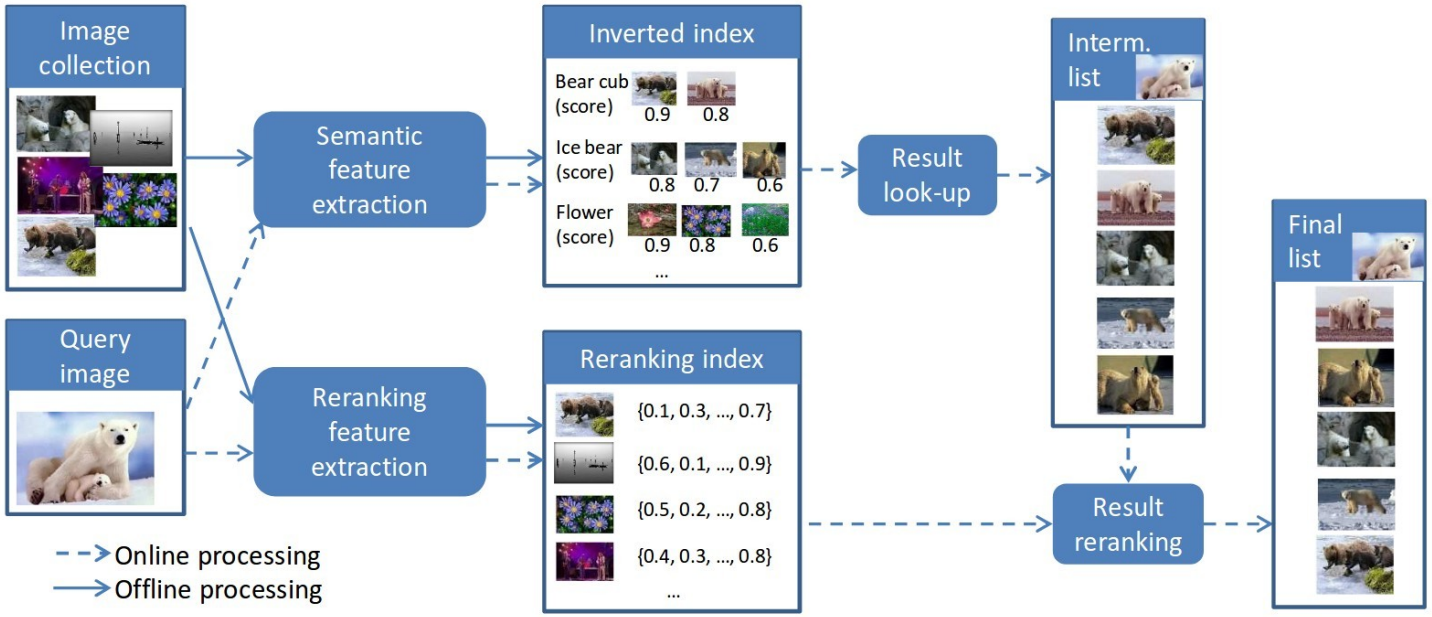
CNN example – VGG16



Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2014

Image retrieval with inverted index

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



Towards image understanding

- Semantic segmentation
- Spatial relationships
- Describing scene

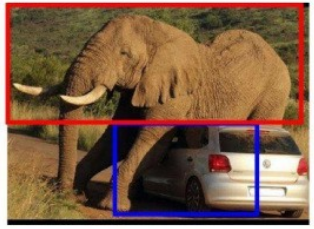
"man in black shirt is playing guitar."

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

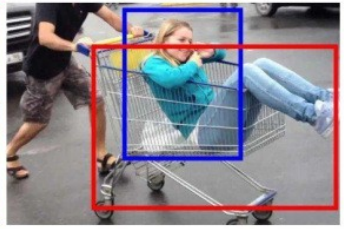
"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

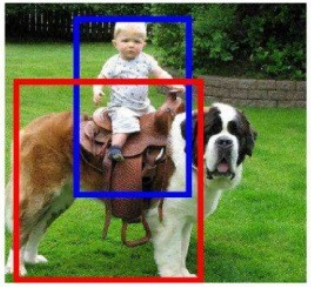
cs.stanford.edu/people/karpathy/deepimagesent/



car under elephant



person in cart

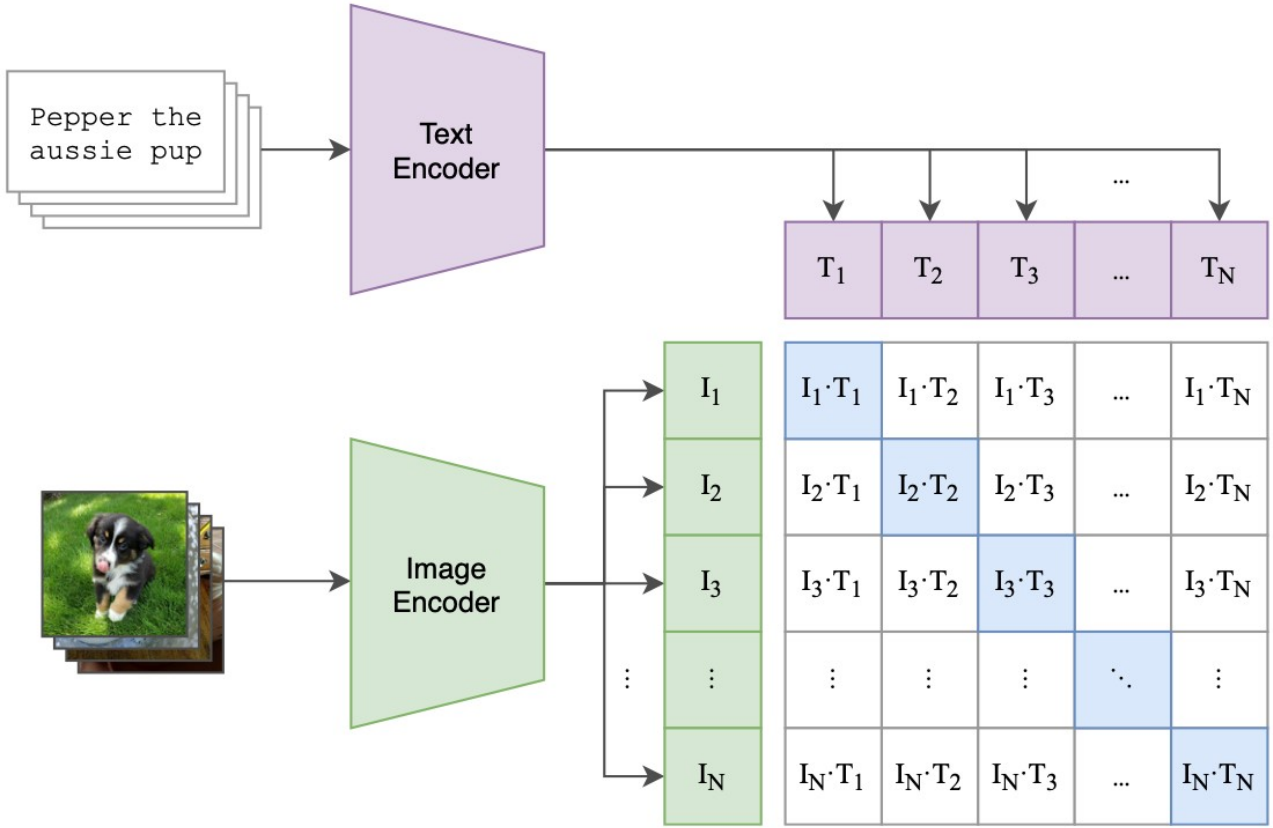


person ride dog

person on top of traffic light

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

Connecting text and images



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML, 2021.

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

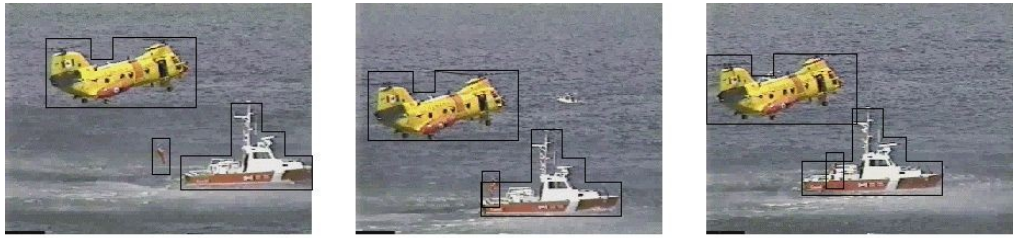
- Efficient access and manipulation of multimedia content
- Complementary to MPEG-4
- Standardized text-less object retrieval
  - D – Object descriptors (audio and video)
  - DS – Description schemes
  - DDL – Description definition language (XML)

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



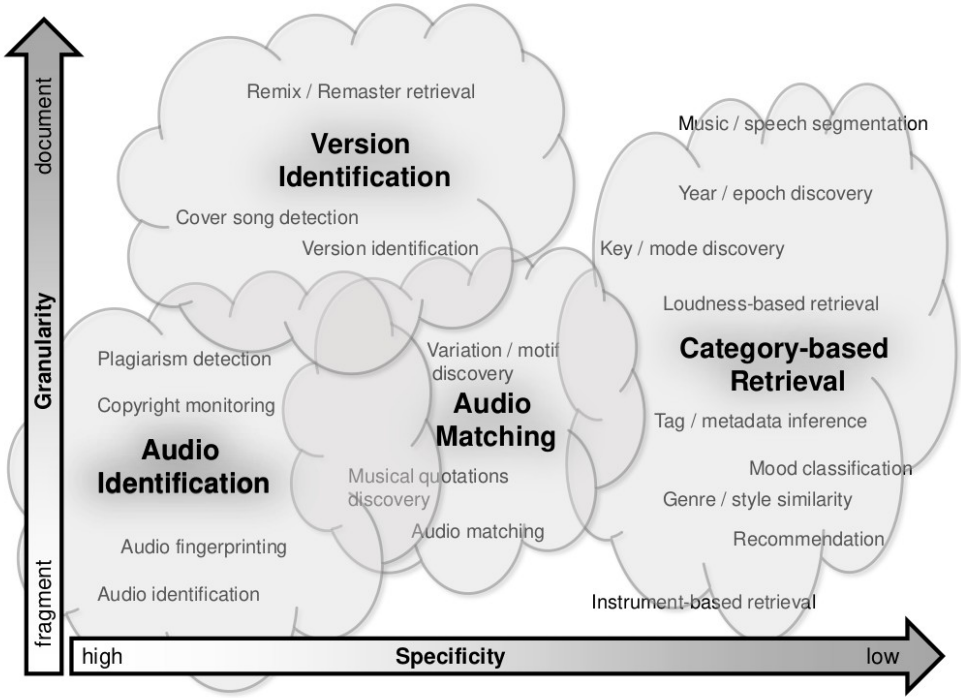
Example: three moving objects, describe relations ...

Three forms of sound

- Speech
  - Words and grammar
  - Can be converted to text
- Music
  - Vocal and/or instrumental sounds
  - Can be represented by a score
- Waveform
  - No dedicated semantic representation
  - Superset

Retrieval in audio

- Identification
  - Exact match
  - Versions, variations
- Segment matching
  - Finding motives, quotations
- Category-based retrieval
  - Genre, mood, tempo
- Recommendation
  - Finding audio with similar qualities

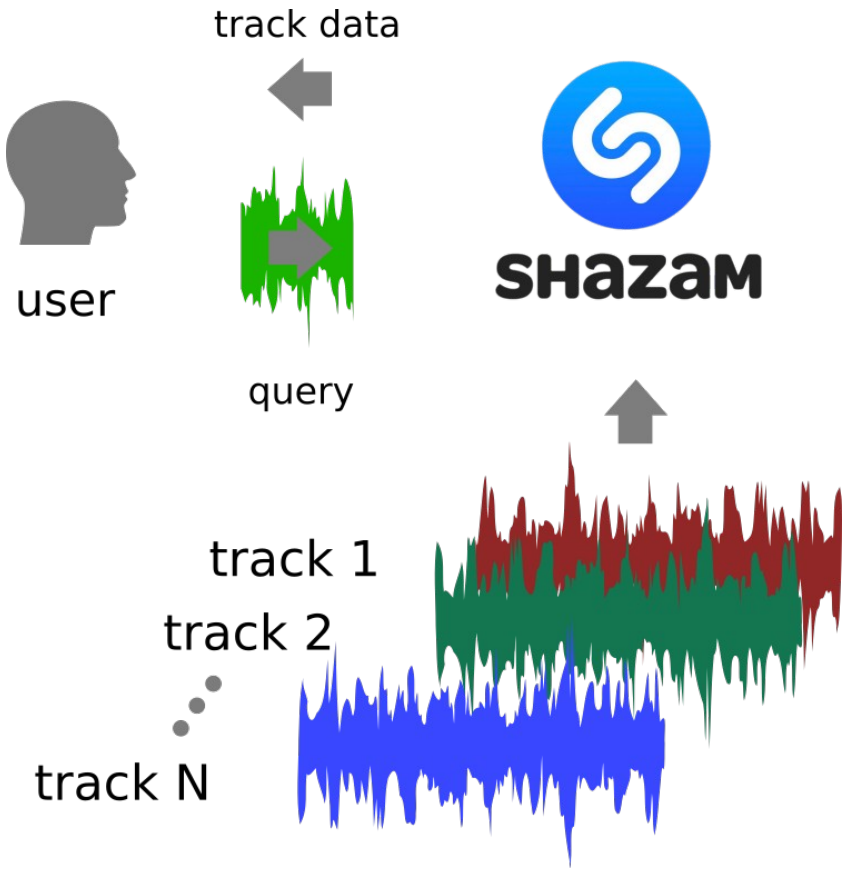


Example-based music search

- Dataset of audio samples (e.g. music)
- Look for most similar sample
- Identification
  - High specificity
- Variations
  - Low specificity
  - Semantic meaning

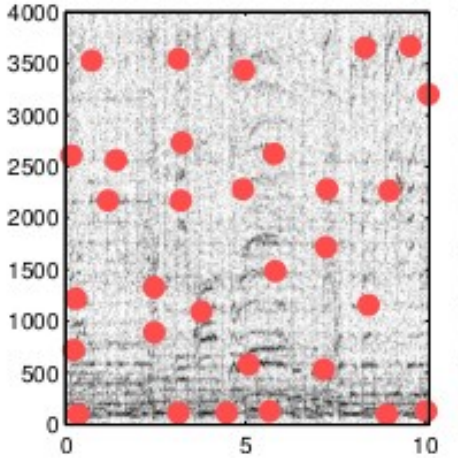
Shazam

- Query by example
  - Short fragments
- Identification
  - High specificity
  - Large database
- Fast, noise resistant
  - Fingerprinting
  - Hashing

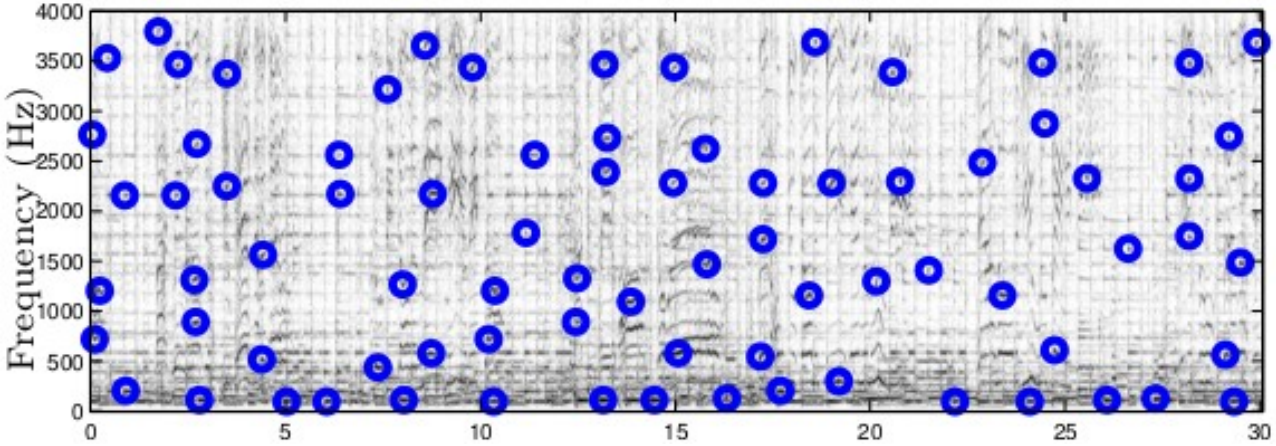


Peak fingerprinting

- Spectrogram
- Local peak strength



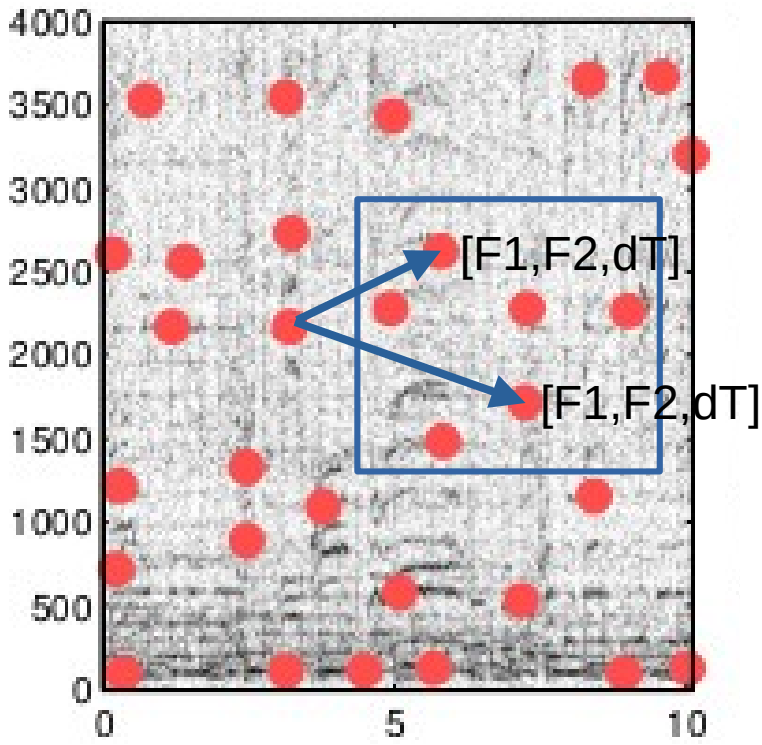
Clip / query



Track / document

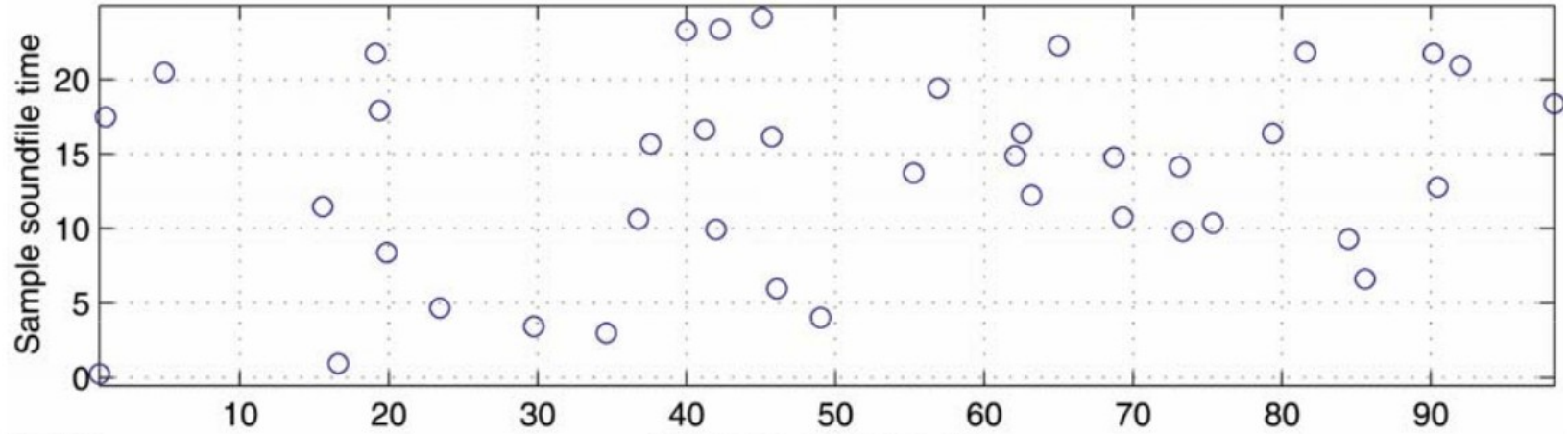
Combinatorial Hashing

- Combinations of peaks
  - Target zone
  - Triplet: F1, F2, dT
  - 30bit hash
- Performance
  - Lower survival (specific)
  - Much higher speed

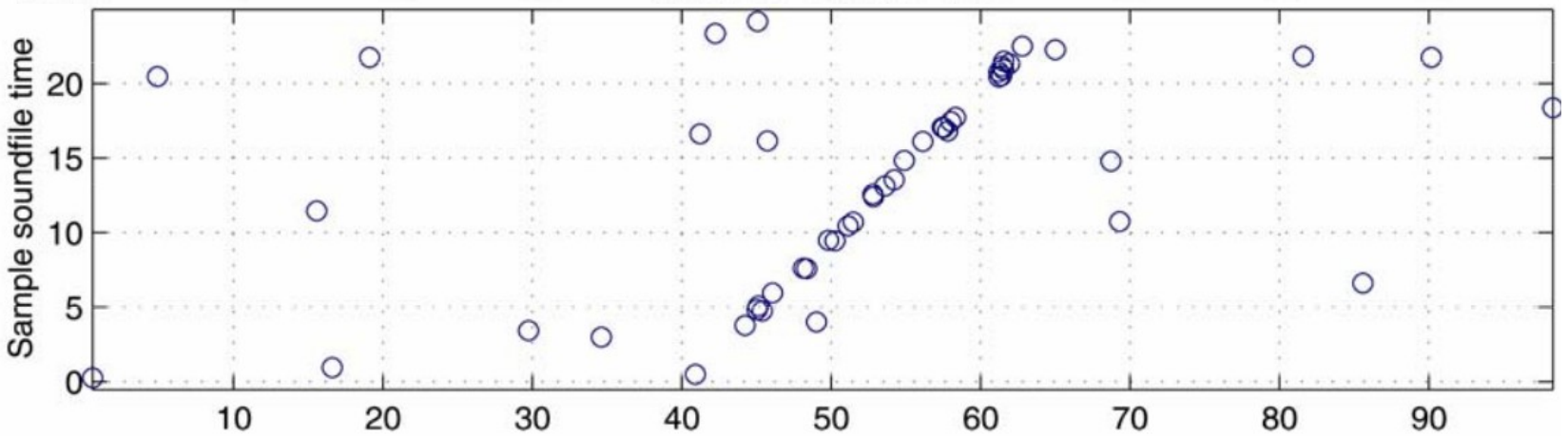


Matching example

Single peak hash



Peak pair hash



Efficient retrieval

- Most descriptors are dense
  - Inverted index not efficient
  - Comparison is slow
- Approximate nearest neighbor
  - Accuracy vs. speed

Approximate nearest neighbor

- Random projection
  - Low-dimensional space
- Structure the space
  - Hierarchical Clustering
  - Product Quantization
  - Hierarchical Navigable Small Worlds
- Locality-sensitive hashing
  - Similar descriptors have the same hash value

Vector databases

- Efficient storage of representations
  - Organization
  - Metadata
- Management
  - Sharding
  - Monitoring
  - Access