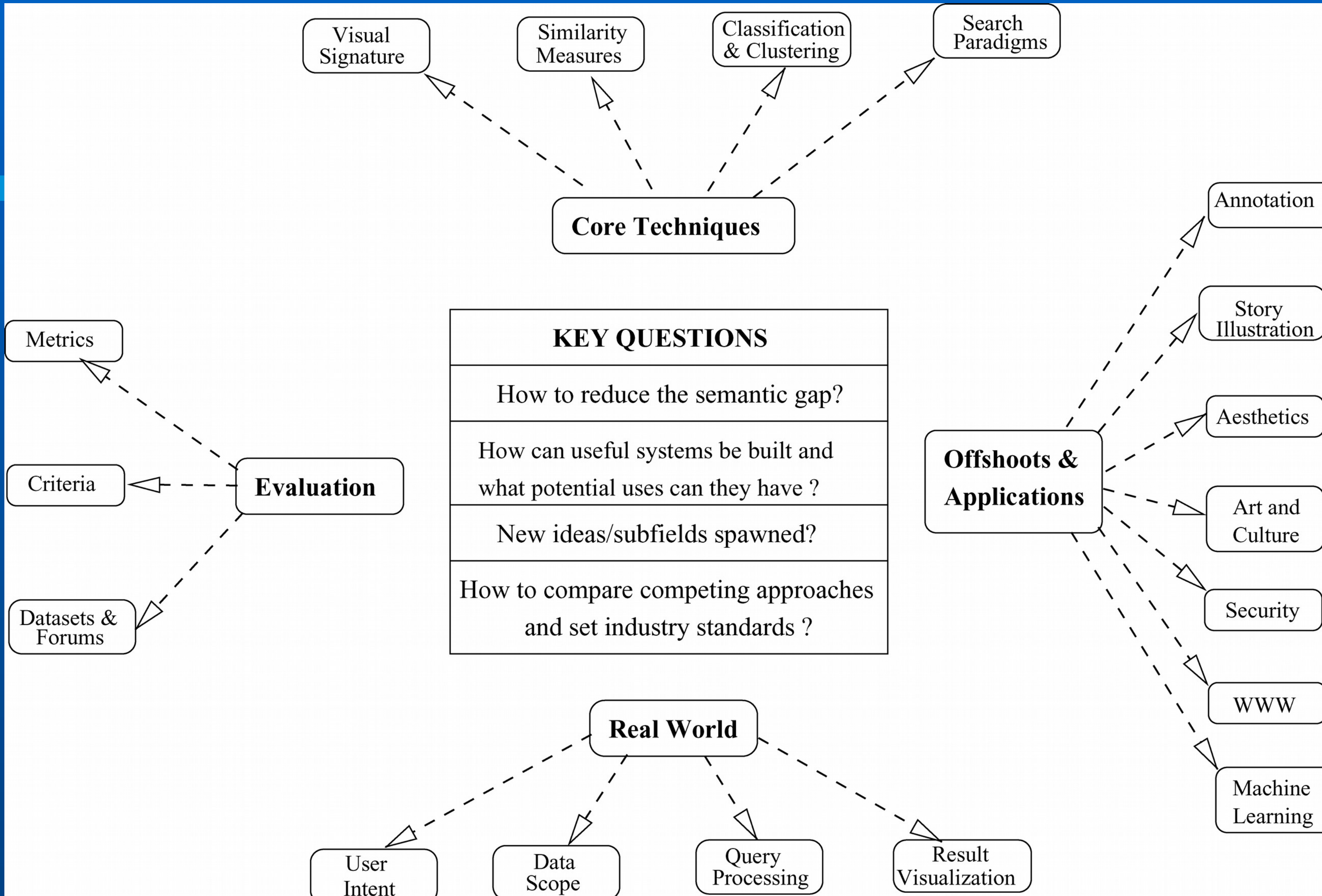
Content-based multimedia retrieval

Georgios Tziritas
Computer Science Department
http://www.csd.uoc.gr/~tziritas

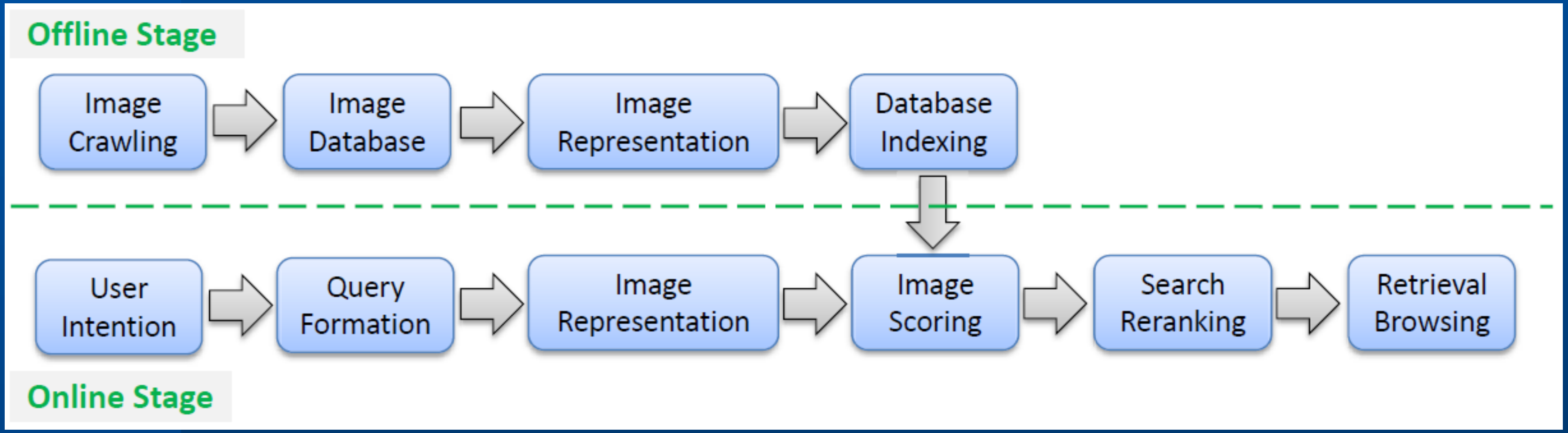

Visual Signature
Similarity Measures
Classification & Clustering
Search Paradigms
Core Techniques
Annotation
Story Illustration
Aesthetics
Art and Culture
Security
WWW
Machine Learning
Offshoots & Applications
Metrics
Criteria
Datasets & Forums
Evaluation
KEY QUESTIONS
How to reduce the semantic gap?
How can useful systems be built and what potential uses can they have ?
New ideas/subfields spawned?
How to compare competing approaches and set industry standards ?
Real World
User Intent
Data Scope
Query Processing
Result Visualization

Content retrieval in real world

User intention : browsing, surfing, classification, search

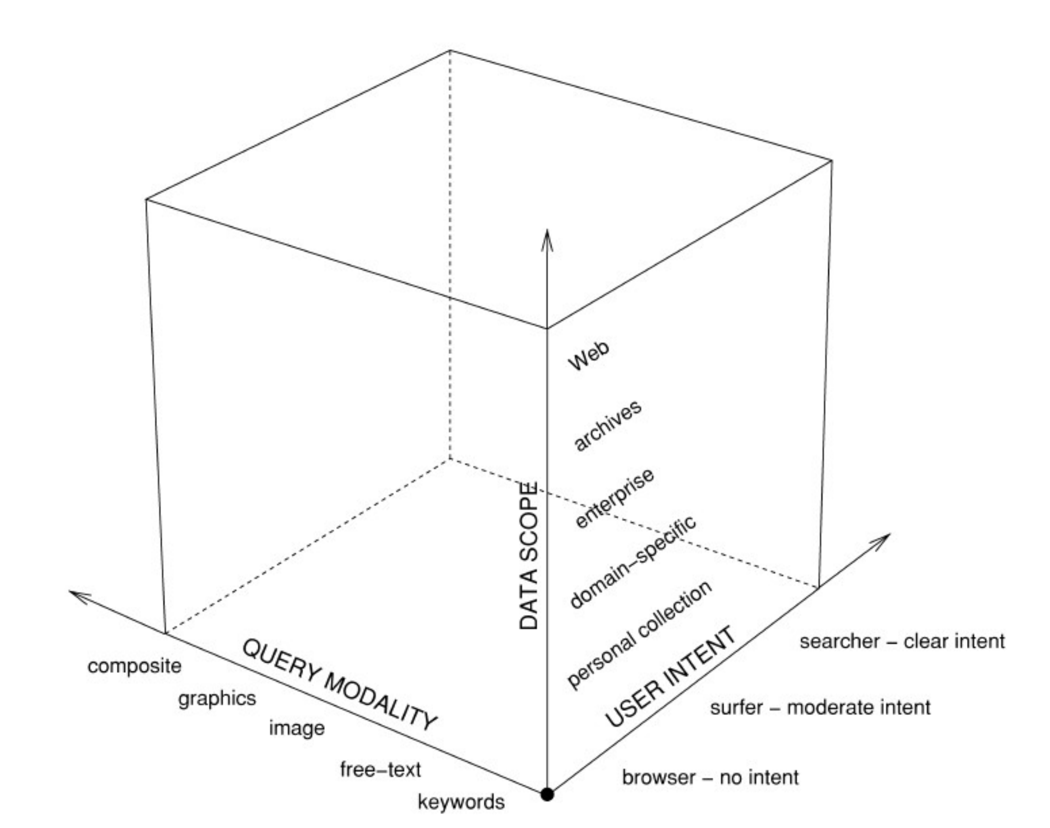
Data scope : personal collection, domain specific, archives, Web

Query formation : key-word, free-text, example, sketch

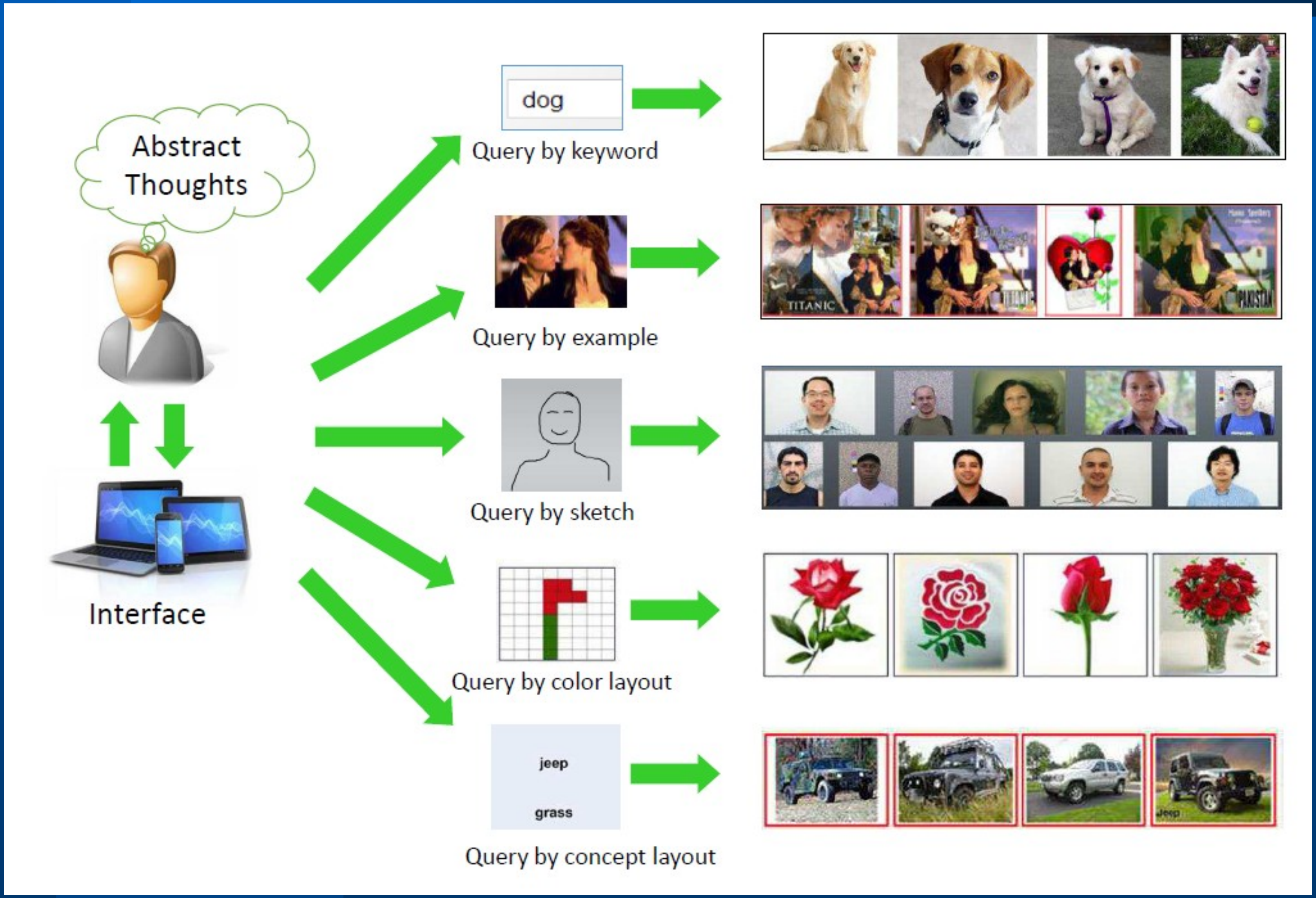


W. Zhou, H. Li and Q. Tian, **Recent Advance in Content-based Image Retrieval: A Literature Survey**, Arxiv, Sept.2017

User / Data / Query



Query modality



It is difficult to precisely express the expected visual content by a query.
The quality of the query has a significant impact on the retrieval results.

W. Zhou, H. Li and Q. Tian, **Recent Advance in Content-based Image Retrieval: A Literature Survey,** Arxiv, Sept.2017

Similarity measures / learning

Agreement with semantics
It is difficulty to describe a high-level semantic concept with low-level visual features
Noise resistance
Computational efficiency
Object scale
Distance properties

Image similarity

$$\begin{aligned} S(\mathcal{X},\mathcal{Y}) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} k(x,y) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \phi(x)^T\phi(y) \\ &= \Psi(\mathcal{X})^T\Psi(\mathcal{Y}). \end{aligned}$$

feature extraction

feature encoding

database indexing

Clustering (hierarchical, grouping, mixtures)

Classification

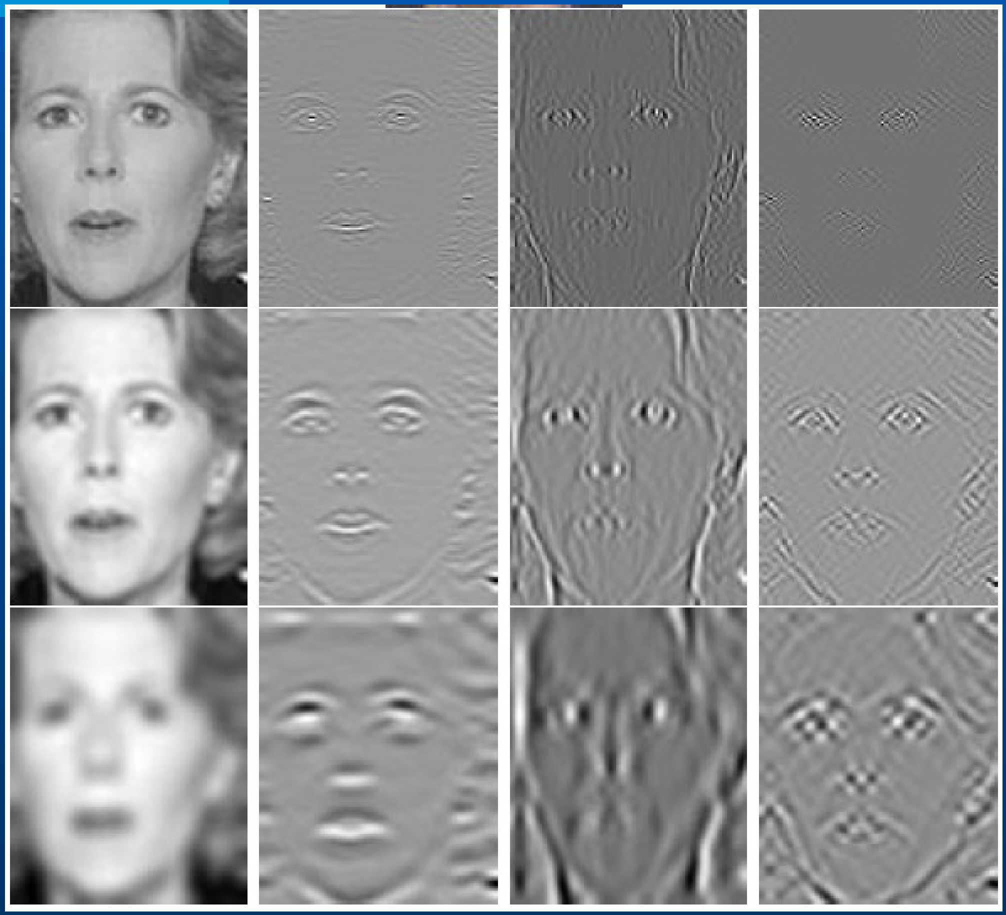
Face detection : color

Skin color detection

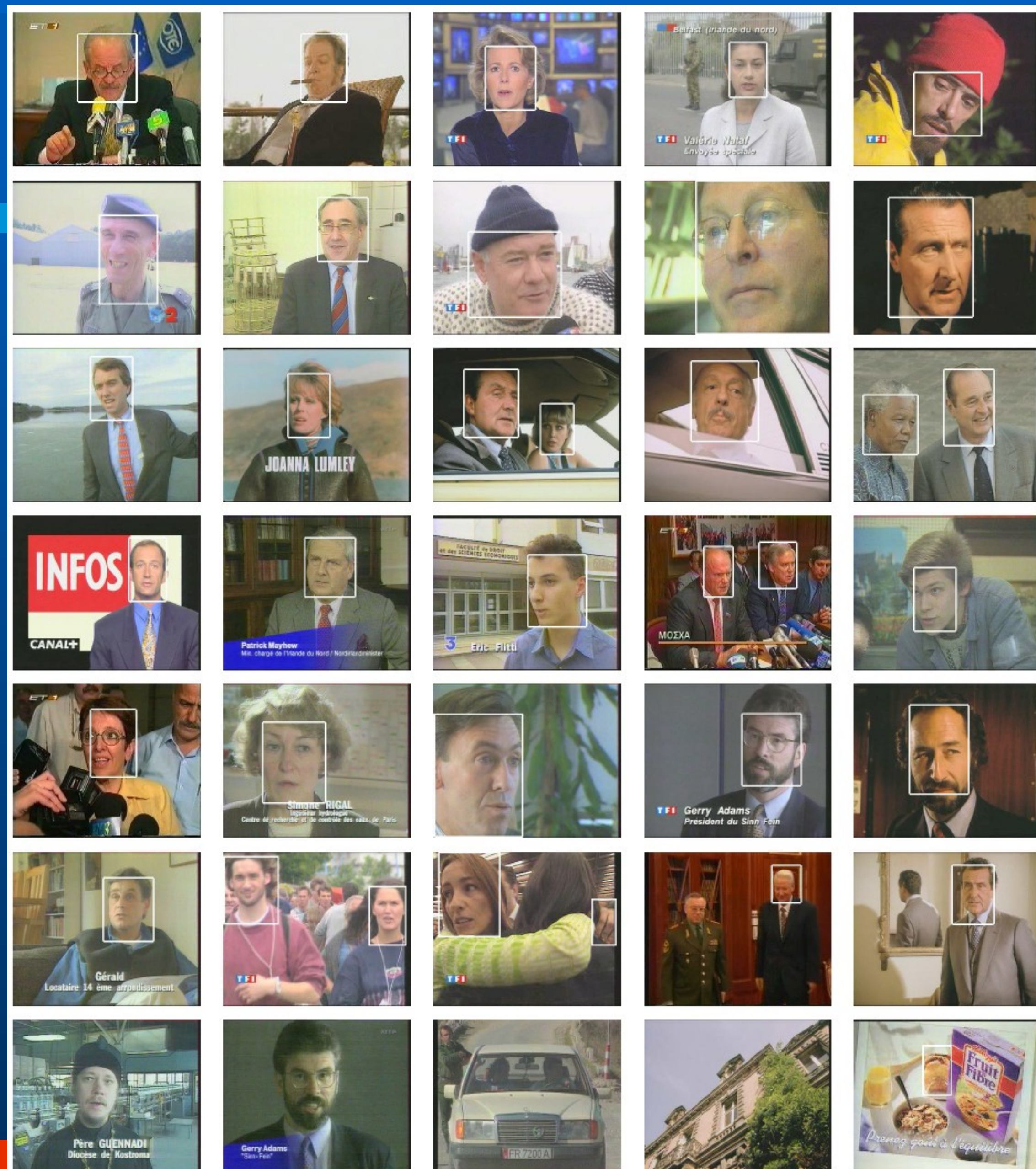
Color system YcbCr, HSV

Face detection : texture

Subband analysis
Discrete Wavelet Frames

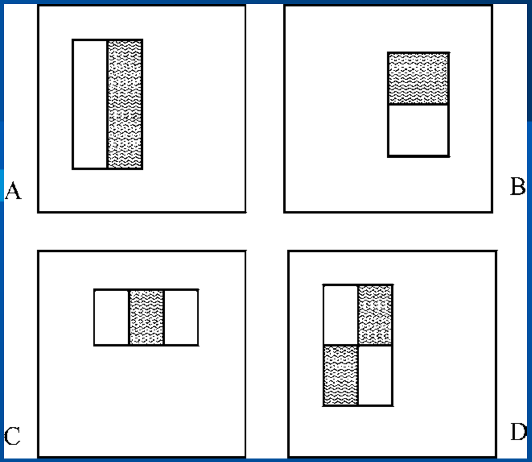


Face detection

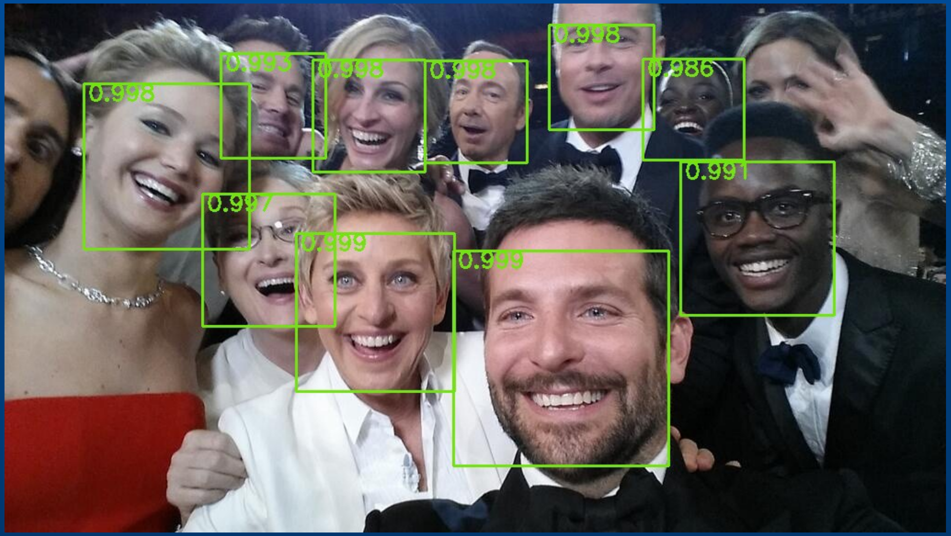


Face detection : learning

Features extraction



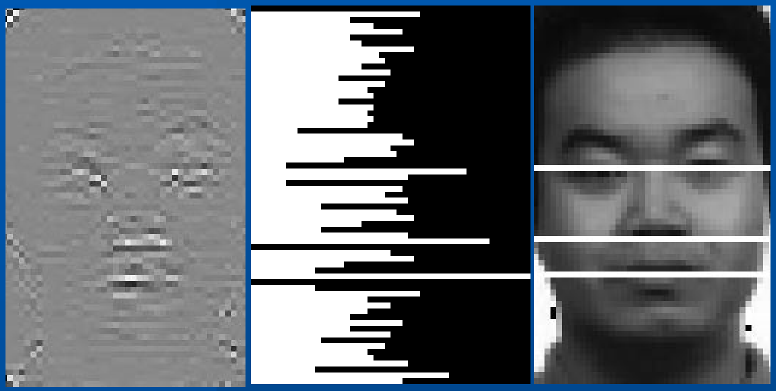
Convolutional neural network



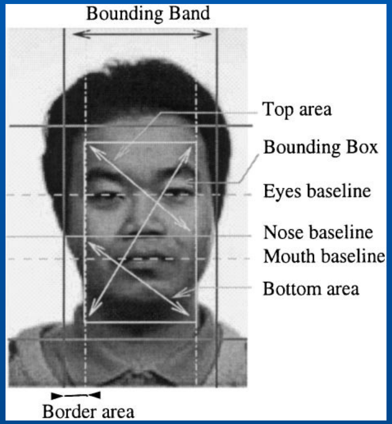
Face recognition

Subband analysis (Discrete Wavelet Transform)

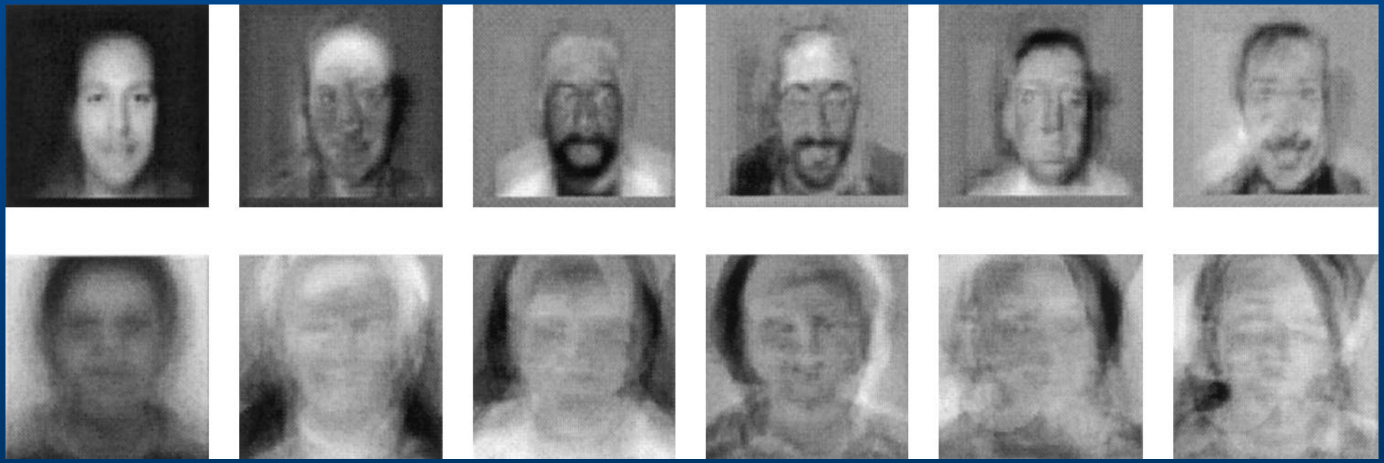
Localisation of characteristic areas



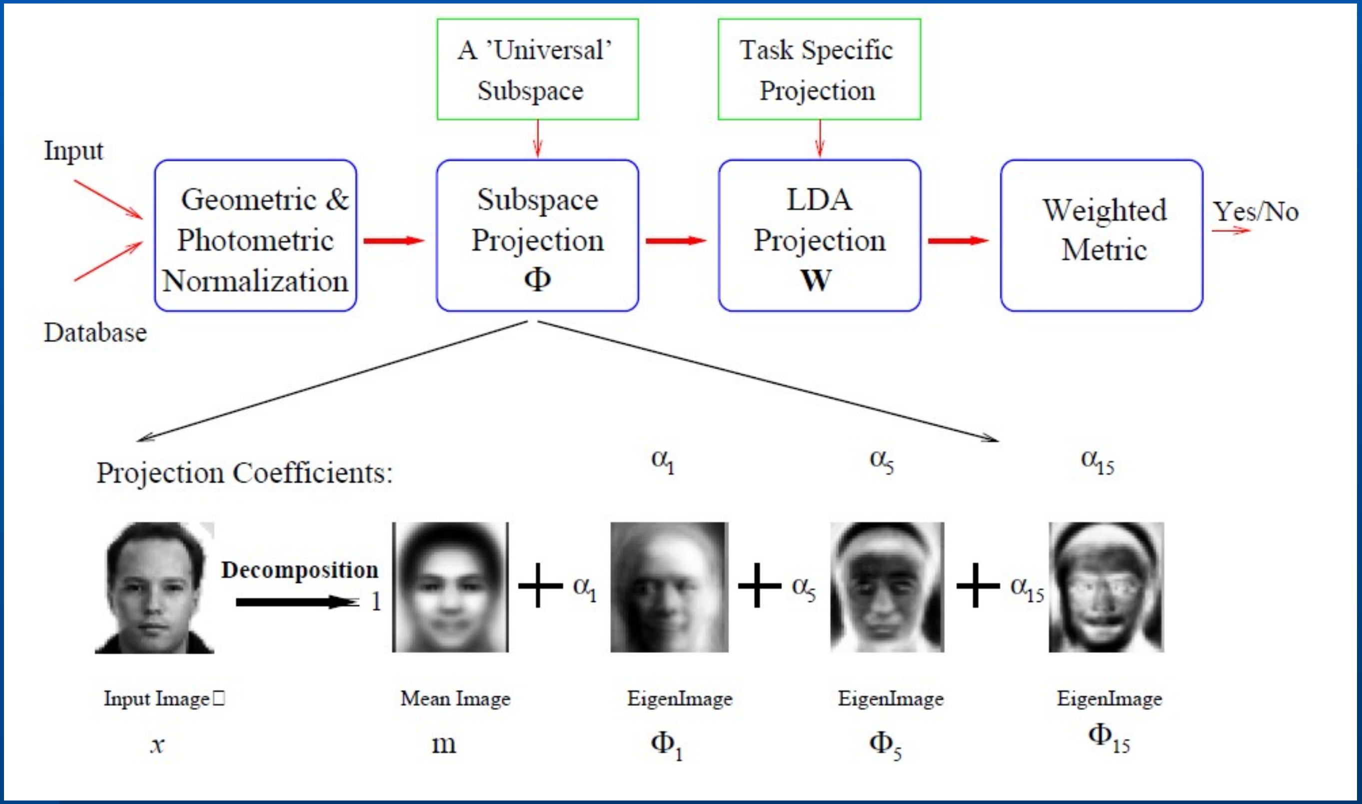
Alignment



Eigenfaces

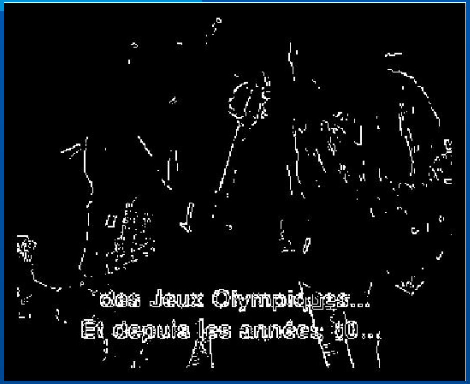


Face recognition

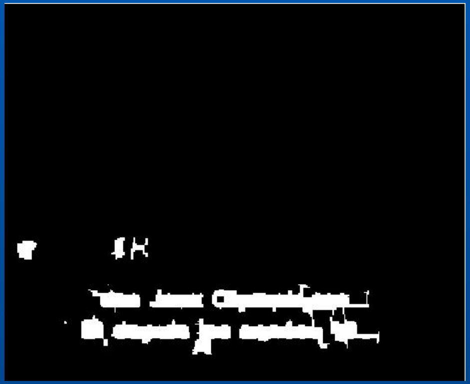


Text detection and recognition

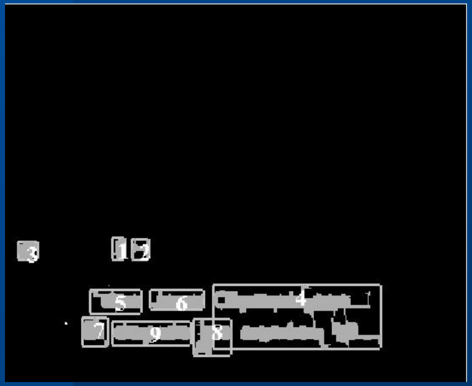
image



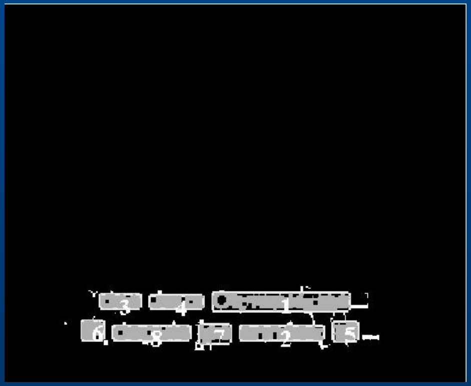
edge detection



background elimination



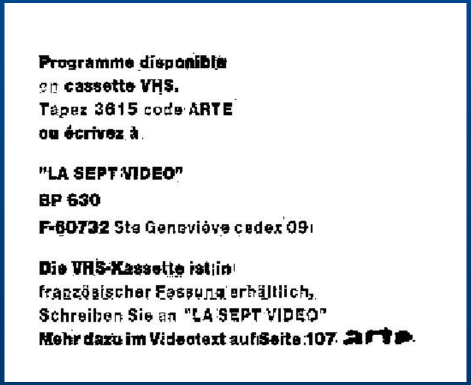
grouping



periodicities

result

Text detection and recognition



Object (class) detection

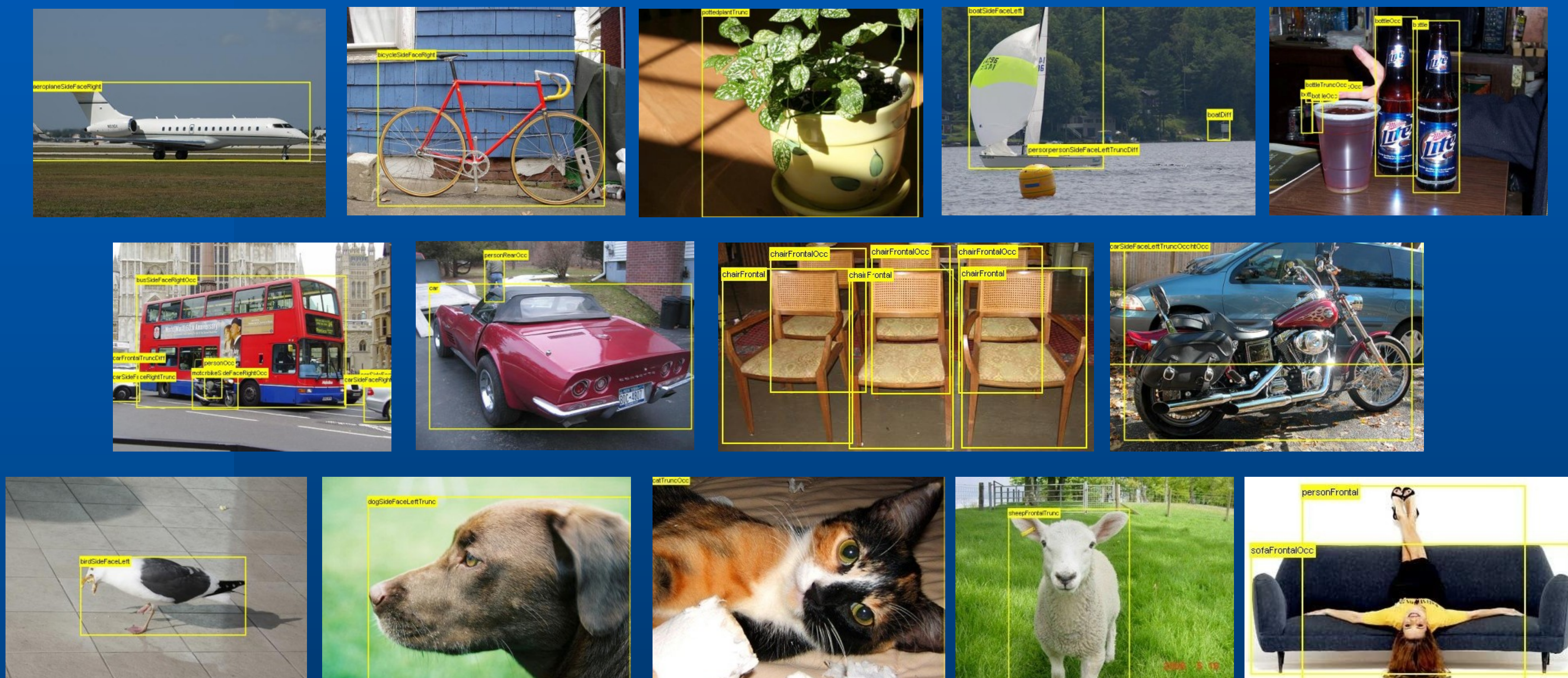
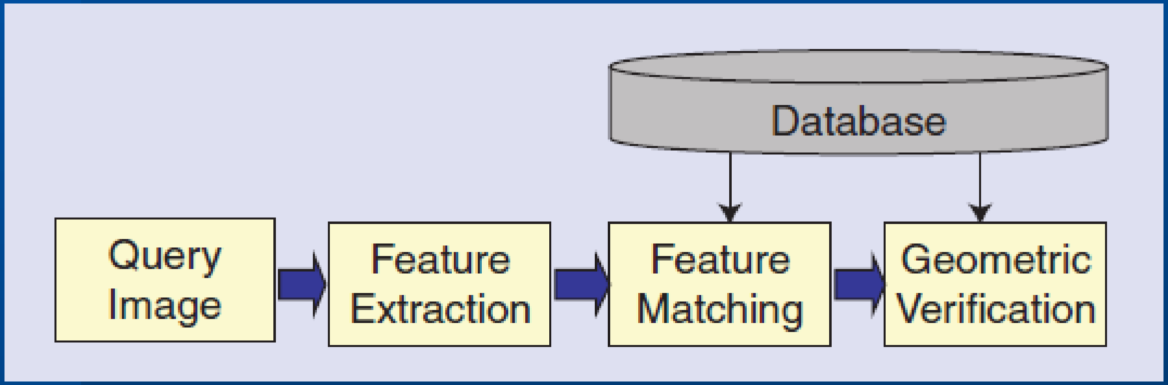


Image analysis / Statistical model / Learning

Content descriptors

Low level content description (color, texture, shape) MPEG-7

Local visual features invariant to geometric transforms



Scale Invariant Feature Transform

s+3 filters

Multiresolution scale space

Key-point detection
Gradient orientation

4σ

$2\sqrt{2}\sigma$

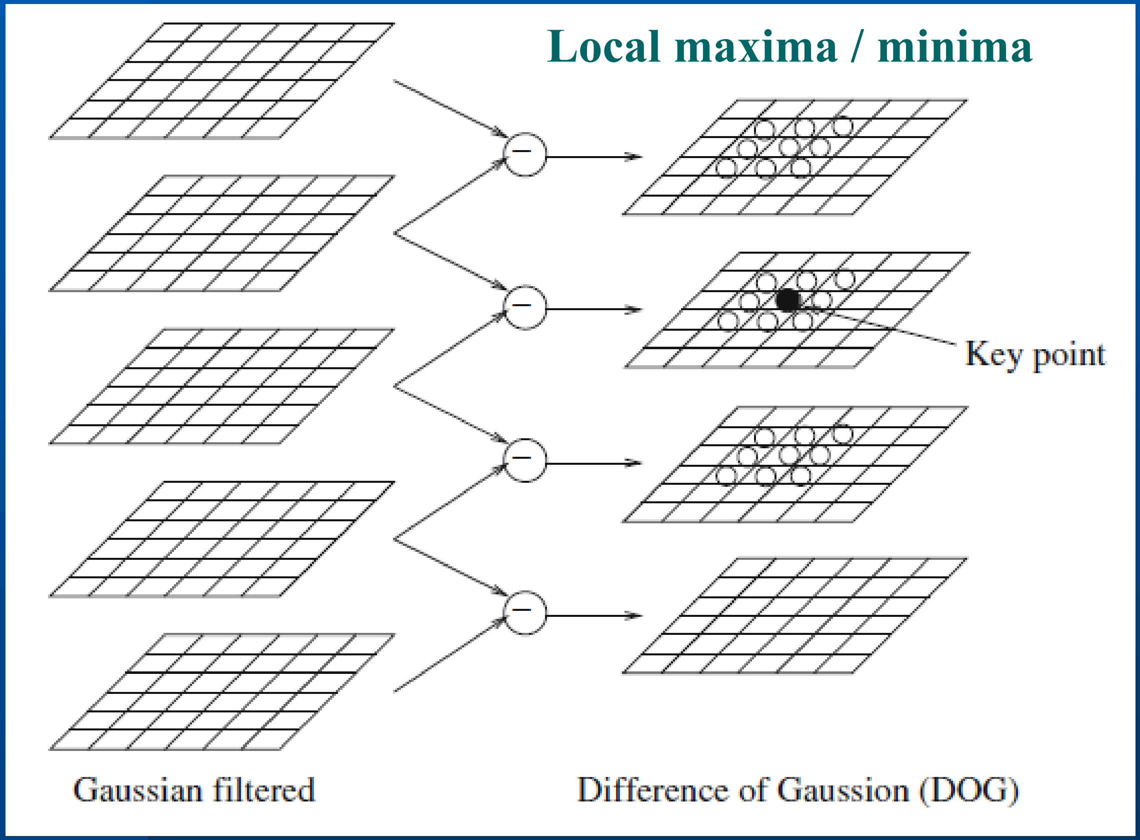
2σ

$\sqrt{2}\sigma$

σ

scale

$\sqrt[s]{2}$



It can well capture the invariance to rotation and scaling transformation and is robust to illumination change

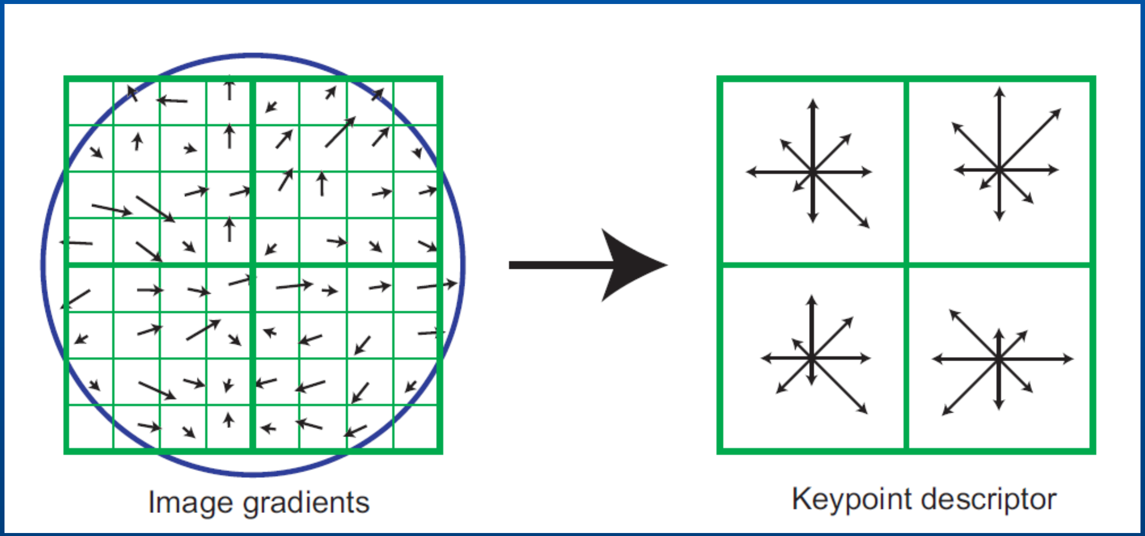
D. Lowe, Distinctive image features from scale-invariant keypoints, *Int. Journal on Computer Vision, 2004*

Key-points

Selection

Scale Invariant Feature Transform

SIFT
Local patch 16x16 pixels
at key-point
Subdivision to 16 sub-blocks
of size 4x4 pixels
Histogram of gradient orientation
(8 bins)
Descriptor 8x4x4=128 dimensions



D. Lowe, Distinctive image features from scale-invariant keypoints, *Int. Journal on Computer Vision*, 2004

Bag of visual words

Compact representation

It may be based on SIFT features (a) for an object or (b) for a video frame

Grouping SIFT features to form object or frame description visual words

Vector quantization for codebook creation (*k-means algorithm*)

Grouping and matching a large number of SIFT descriptors is a computational challenge

The quantization result of a single local feature can be regarded as a high-dimensional binary vector, where the non-zero dimension corresponds to the quantized visual word.

Visual words are rich in encapsulation of basic visual characteristics, despite the inevitable uncertainty

With small codebook and feature space coarsely partitioned, irrelevant features with large distance may also fall into the same cell.

When the codebook size is large which means the feature space is finely partitioned, features proximate to the partition boundary are likely to fall into different cells.

Image retrieval : learning

Neural network

Learning by training

- content representation
- similarity criterion (classification)

Bridging the semantic gap

Pretraining in large base / adaptation to specific classes

Indexing

Image index refers to a database organizing structure to assist for efficient retrieval of the target images.

Inverted file indexing

Sparse matrix : Rows correspondent to images and columns denote visual words
Each visual word is followed by an inverted file list of entries.
In on-line retrieval, only those images sharing common visual words with the query image need to be checked.
Thus, the number of candidate images to be compared is greatly reduced.

Hashing

Partition the feature space, so that similar images can be found in nearby areas
The large dimension features are encoded into low-dimension binary codes for search by similarity
Semantically similar data must have close binary codes

Geometric verification

By including contextual information, the discriminative capability of visual codebook can be greatly enhanced.
Loose spatial consistency from some spatially nearest neighbors can be imposed to filter false visual-word matches

B. Girod et al., Mobile visual search, IEEE Signal Processing Magazine, 2011

Image scoring

Feature distance

$$D(I_q, I_m) = \left(\sum_{i=1}^{N} |q_i - m_i|^p \right)^{\frac{1}{p}}$$

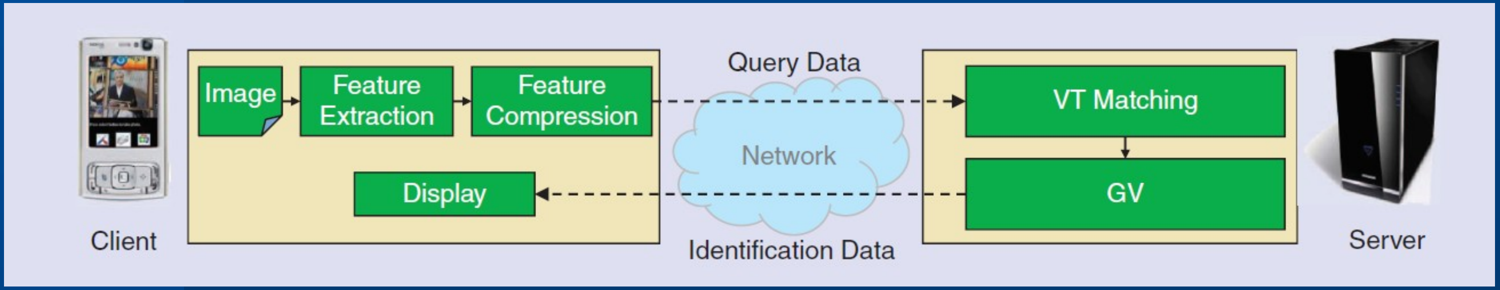
BoVW : weighted visual word histogram

Weighting visual words

Term frequency in the document

Inverse document frequency $\log\left(\frac{N}{nt}\right)$

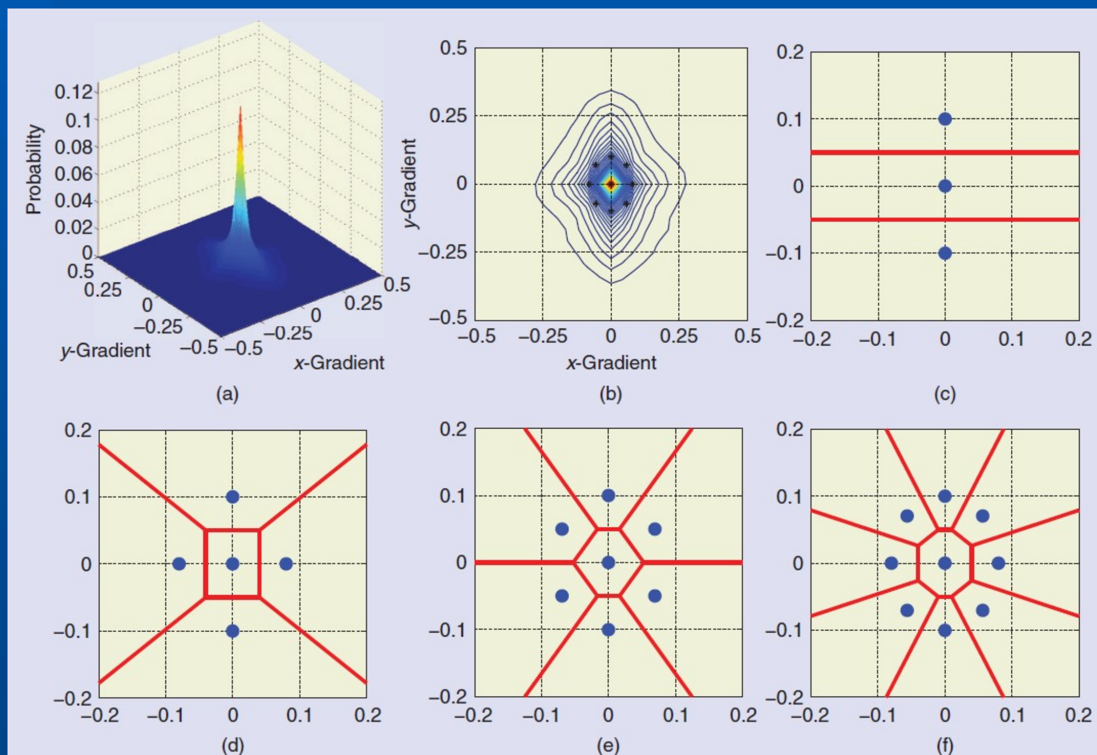
Mobile visual search



B. Girod et al., Mobile visual search, *IEEE Signal Processing Magazine*, 2011

Compact feature description : compressed histogram of gradient

Gradient at key-points

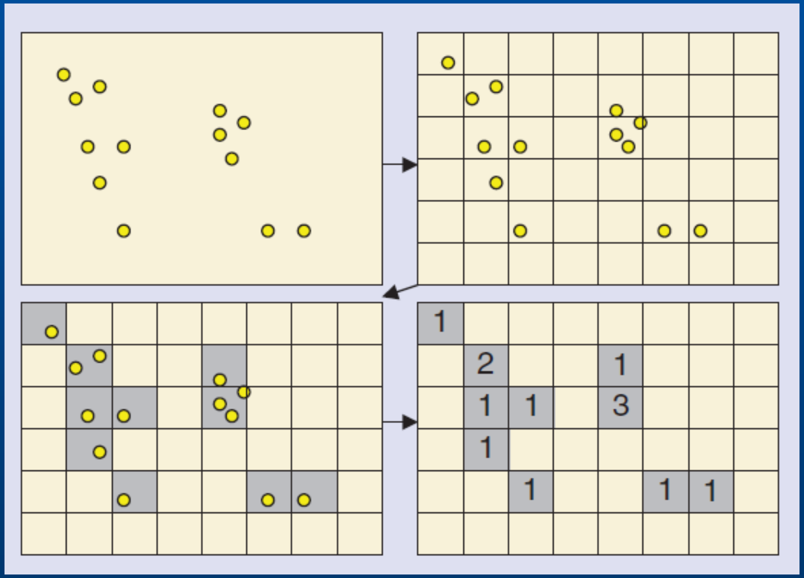


VQ constellations

B. Girod et al., Mobile visual search, IEEE Signal Processing Magazine, 2011

Compact feature description : location histogram

The interest points are spatially clustered
It is possible to compress feature location data efficiently

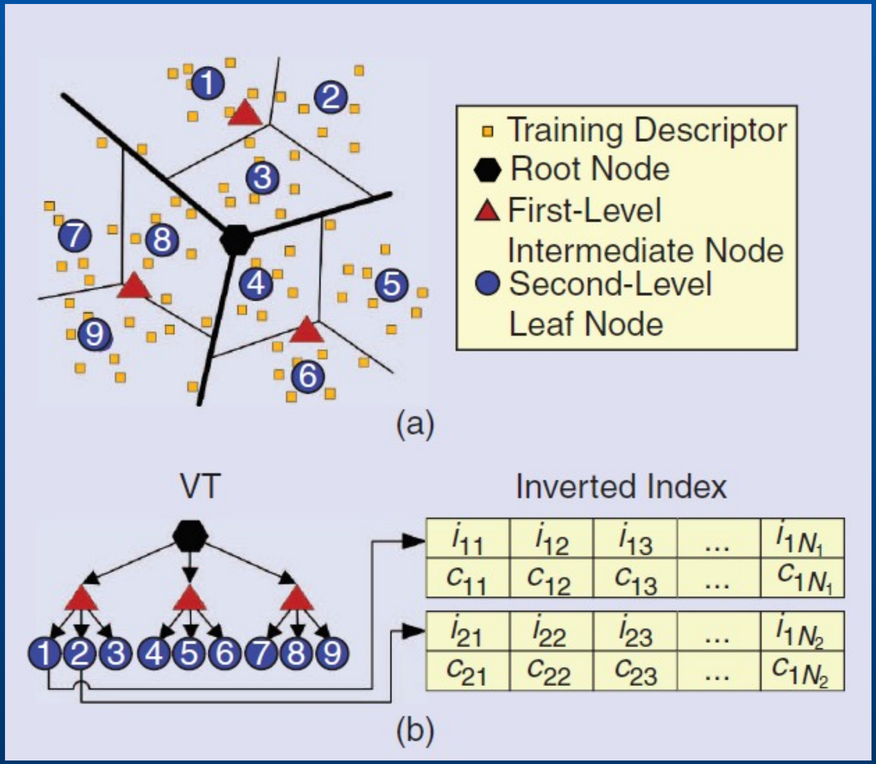


B. Girod et al., Mobile visual search, IEEE Signal Processing Magazine, 2011

Indexing / query

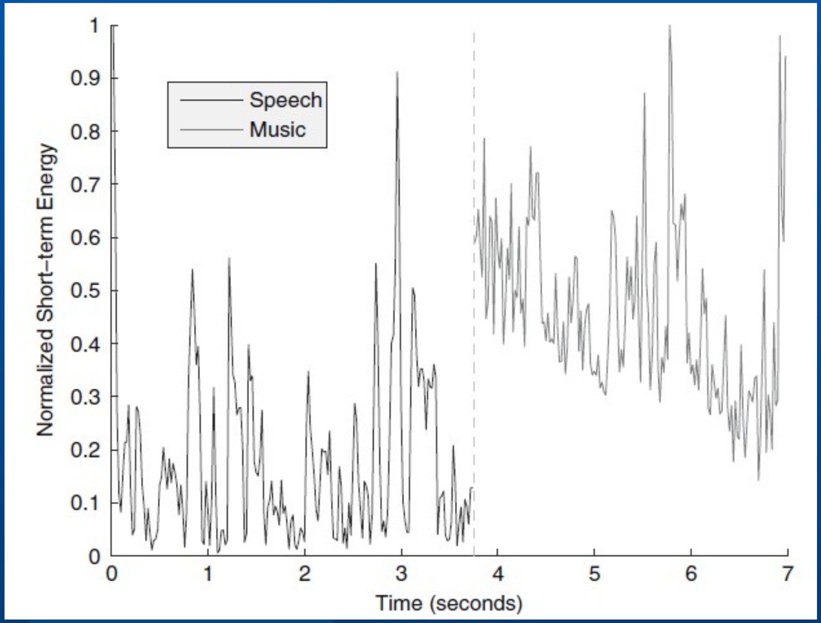
Vocabulary hierarchical tree

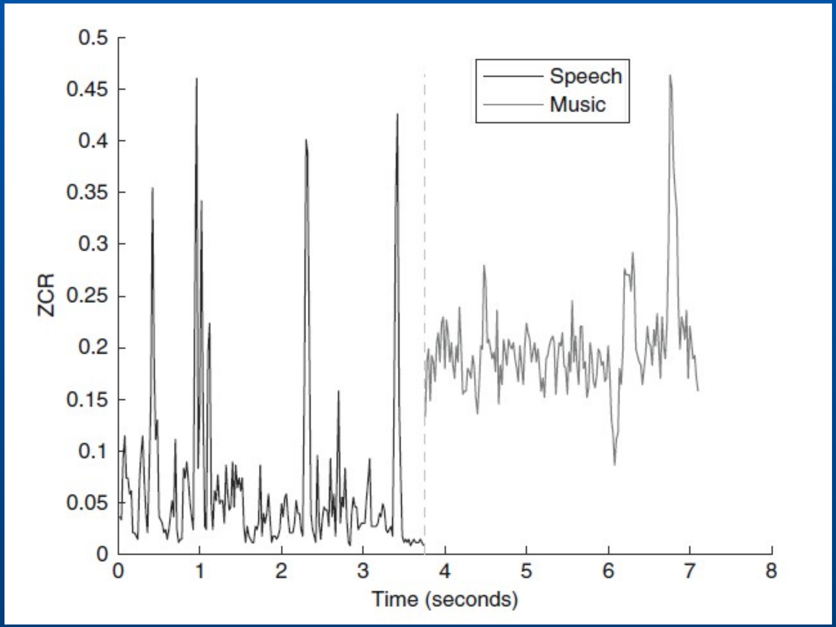
An approximate nearest neighbor search is achieved by propagating the query feature vector from the root node down the tree by comparing the corresponding child nodes and choosing the closest one.



B. Girod et al., Mobile visual search, IEEE Signal Processing Magazine, 2011

Speech / music discrimination





Feature extracted on frames : mean over standard deviation