Tecniche di indicizzazione automatica

Automatic Indexing

- **Overview**
  - Text
  - Speech
  - Images
  - Moving pictures (videos)

Indexing text

- **The indexing process associates (weighted) index terms to documents**

- **Index terms can be**
  - Words chosen from a controlled vocabulary
  - Words automatically extracted
  - Steams (e.g. print-)
  - Noun phrases automatically extracted
  - Other metadata

Indexing text

- **Experience has shown that using weighted single terms offers the best performance**
  - Of course that depends crucially on the choice of the term-weighting system
- **Document search is performed by searching for index terms**
  - Documents associated with qualifying index terms are retrieved
  - Documents are ranked according to weights of index terms

Indexing text

- **The indexing process produces an incidence matrix:**

	d_1	…	d_i	…	d_m
t_1	w_{11}	…	w_{1i}	…	w_{1m}
…	…	…	…	…	…
t_k	…	…	w_{ki}	…	…
…	…	…	…	…	…
t_n	w_{n1}	…	w_{ni}	…	w_{nm}

Indexing text

- **Models to assess document relevance:**
  - Boolean model
  - Fuzzy logic model
  - Vector space model
  - …

Boolean model

- **A query may contain logical operator and/or**
  - The query “digital and library” retrieves documents associated with both terms
  - The query “digital or library” retrieves documents associated with at least one of the two terms
- **Boolean logic is used to process more complex queries**

Fuzzy logic model

- **Extends the Boolean model in such a way that also weights are considered to assign a score to retrieved documents**

- **Suppose that term t_1 and t_2 have weight w_1 and w_2 in document d**

- ***d* has score:**
  - $min\{w_1,w_2\}$ for query t_1 and t_2
  - $max\{w_1,w_2\}$ for query t_1 or t_2

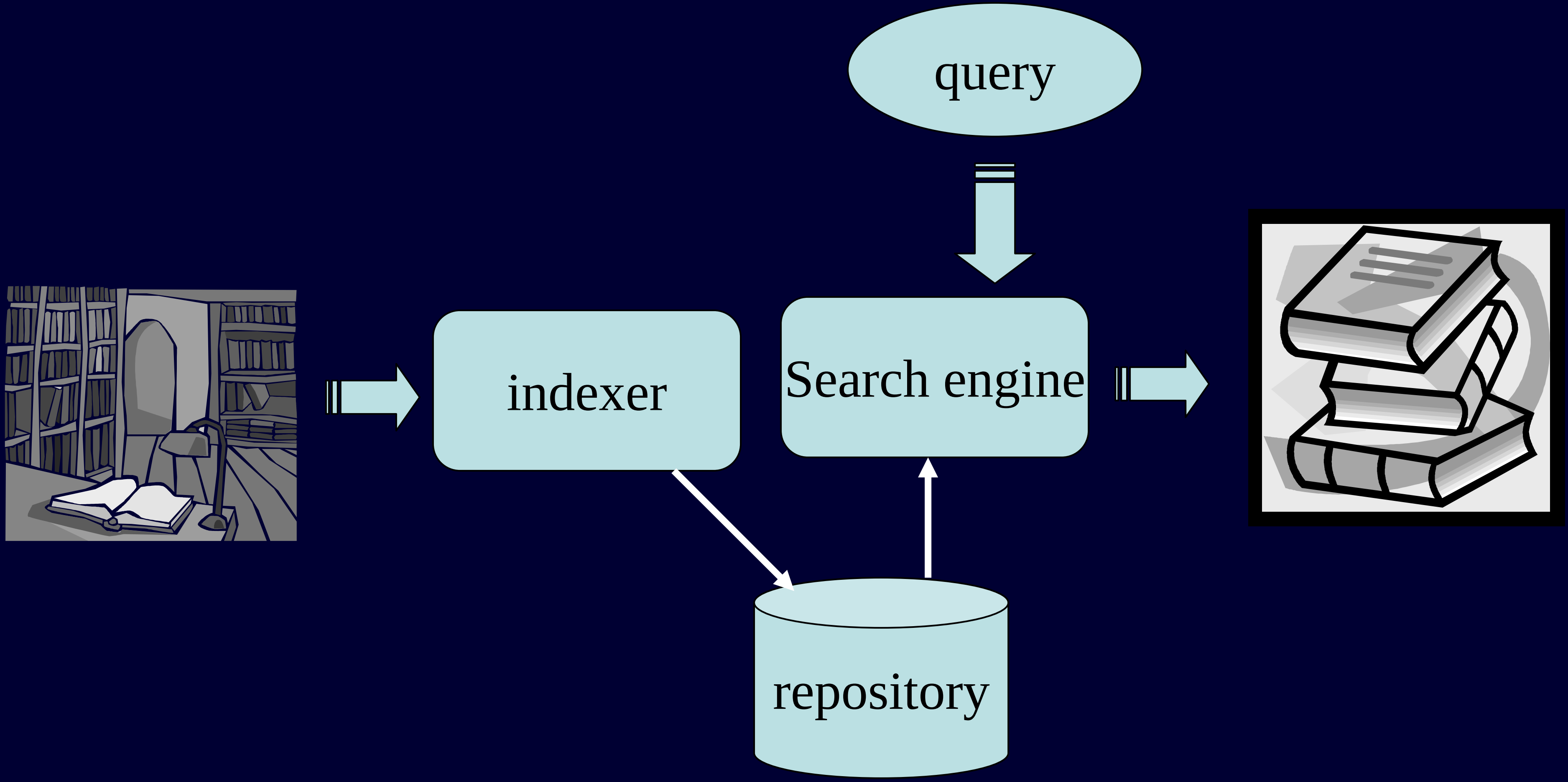
Vector space model

- Documents and queries can be viewed as vectors of of weights (each term is a dimension)
- The score is the distance between a query (vector) and the documents (vectors)

Automatic extraction of weighted index terms

- **A widely used technique is the *tfidf* weighting function (term frequency inverse document frequency):**
  - The more frequently a term appear in a document the more significant it is for that document: term frequency (*tf*)
  - The more frequently a term occur in the entire collection the less selective it is: document frequency (*df*)
- **The weight is directly proportional to the *tf* and inversely proportional to the *df* (*idf*)**

Text documents:Overall view



query

indexer

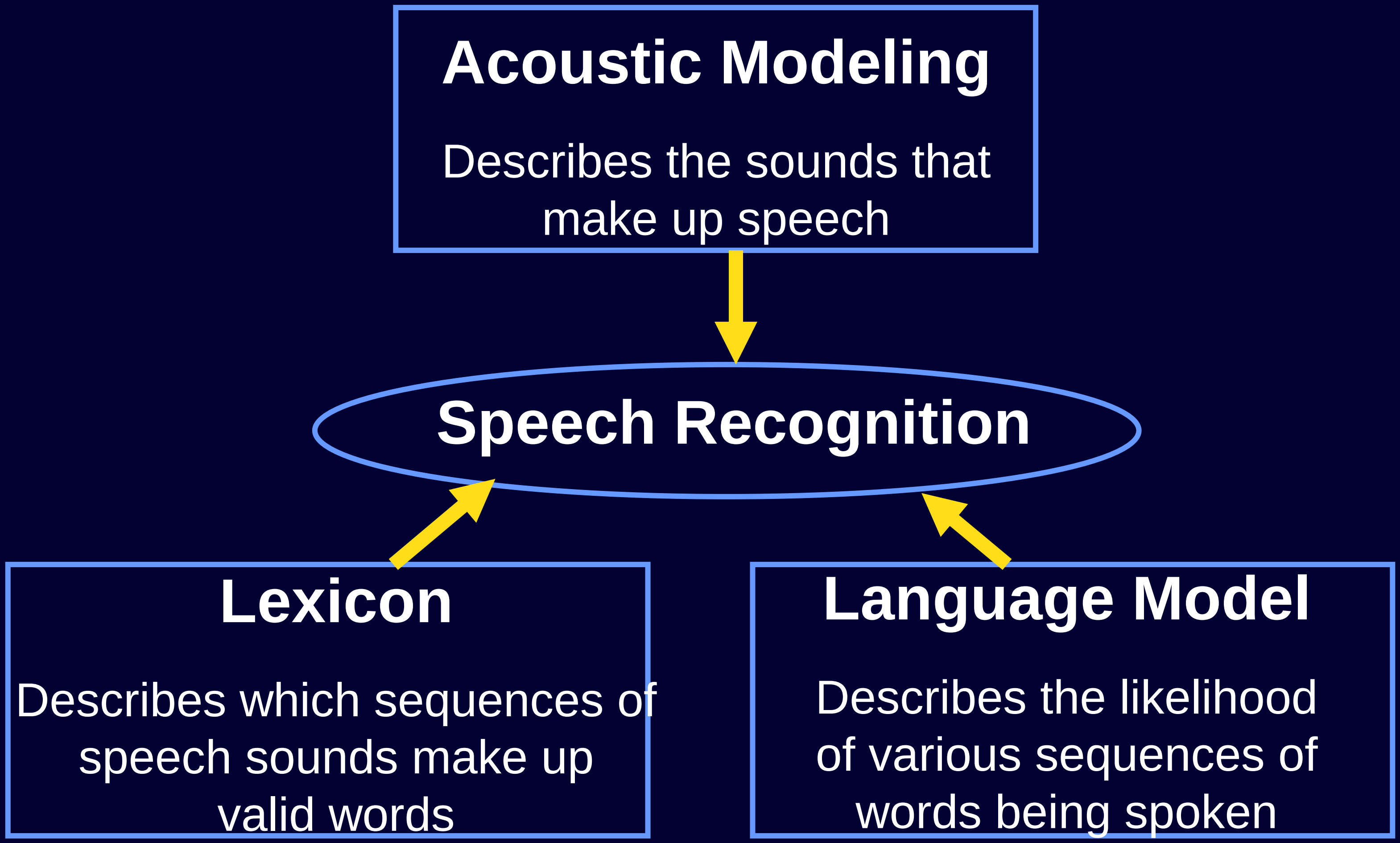
Search engine

repository

Indexing speech

- Generates transcript to enable text-based retrieval from spoken language documents
- Improves text synchronization to audio/video in presence of scripts
- Supplies information necessary for library segmentation and multimedia abstractions
- Provides speech interface to digital library

Indexing speech



Acoustic Modeling

Describes the sounds that make up speech

Speech Recognition

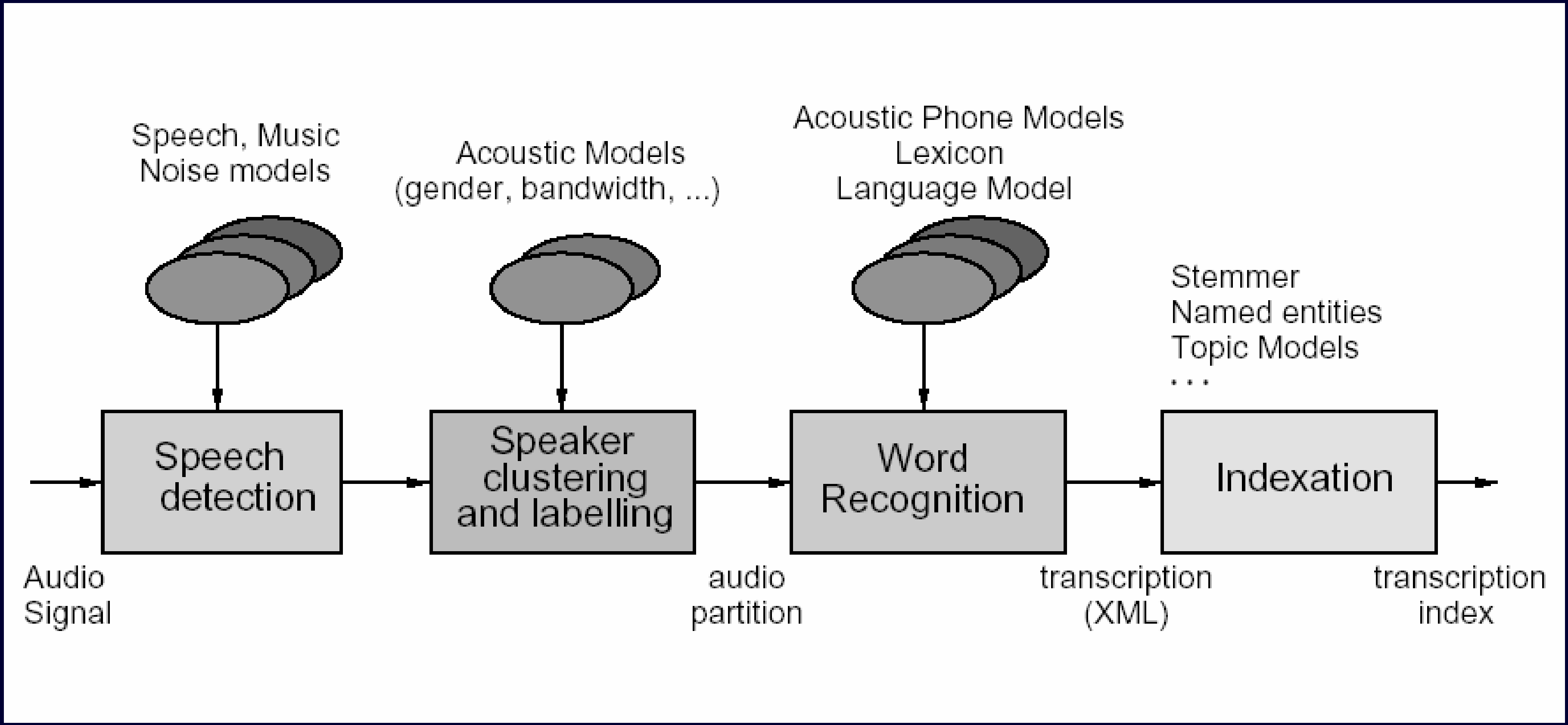
Lexicon

Describes which sequences of speech sounds make up valid words

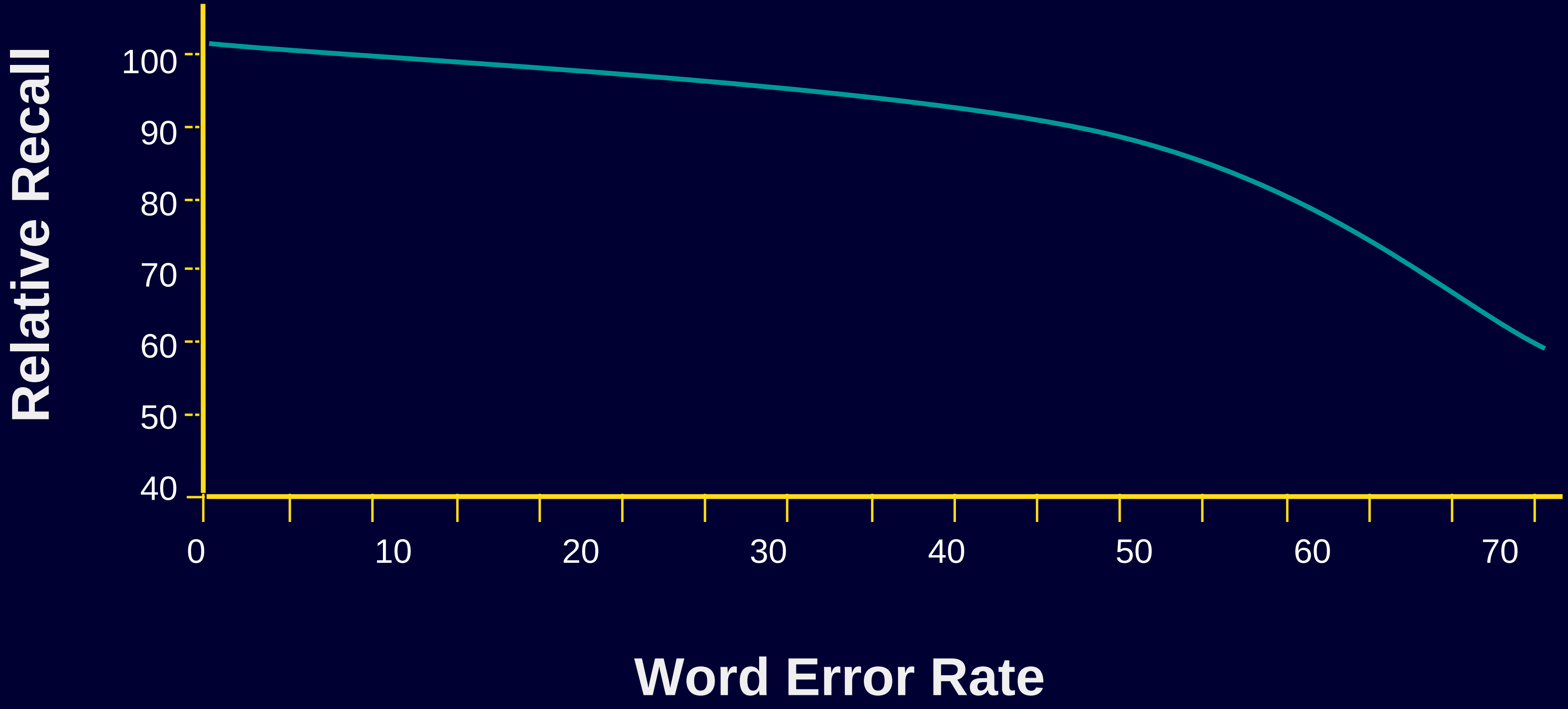
Language Model

Describes the likelihood of various sequences of words being spoken

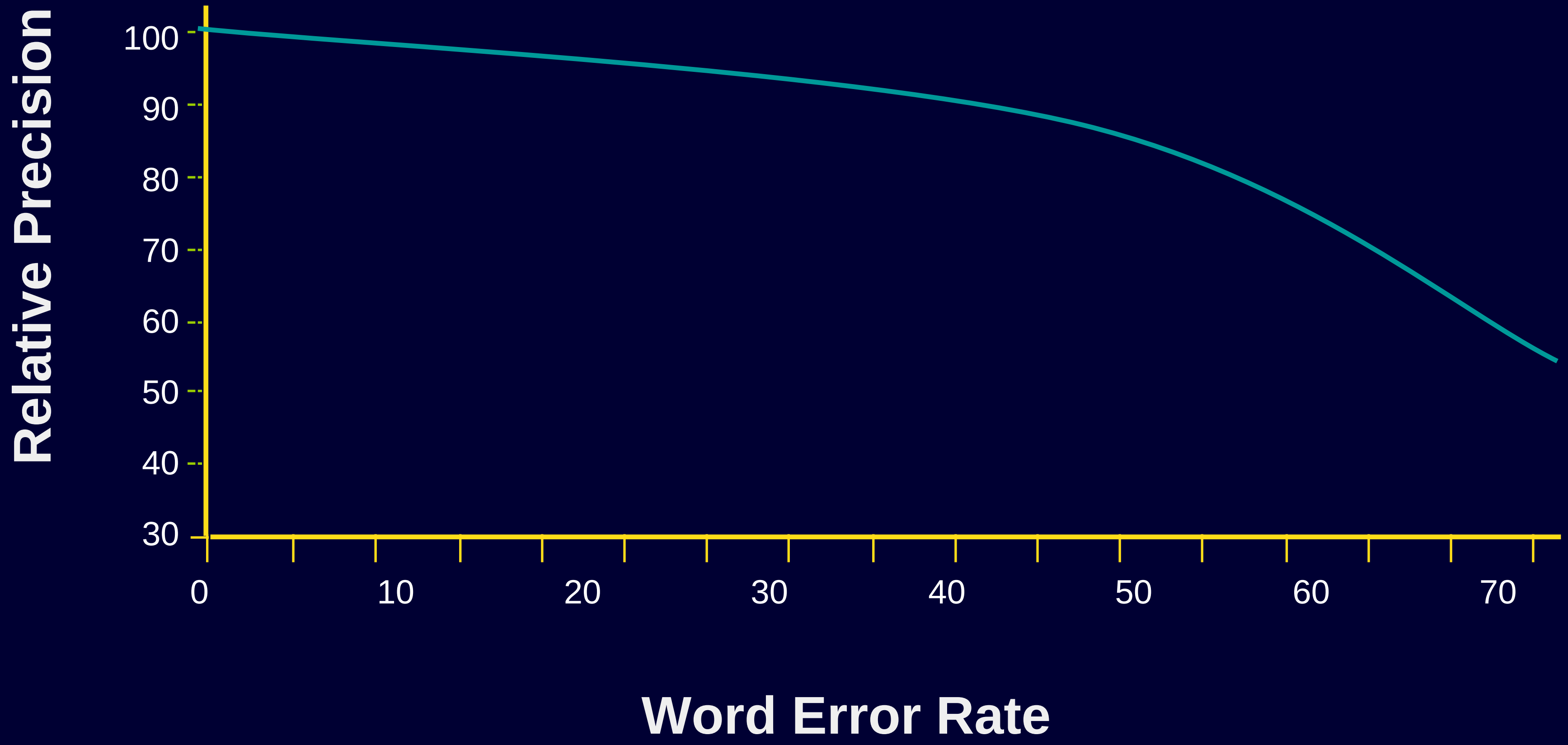
Indexing speech



Text retrieval precision vs. Speech accuracy



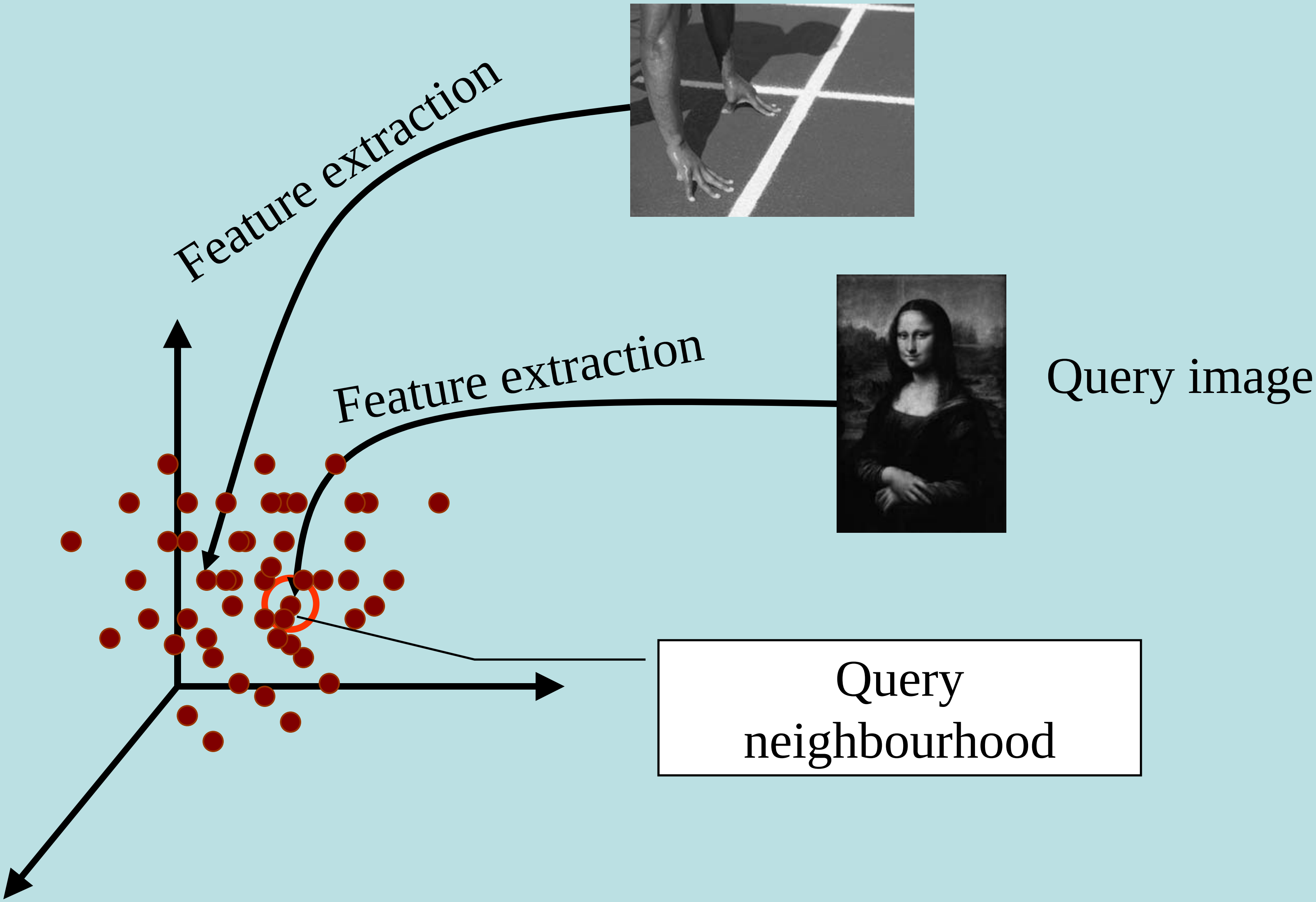
Text retrieval precision vs. Speech accuracy



Indexing images

- **The automatic indexing process associates images with features describing their physical content**
  - Colour
  - Textures
  - Shapes
  - Spatial organisation
- **Image search is performed by using feature similarity**

Similarity search



Feature extraction

Feature extraction

Query image

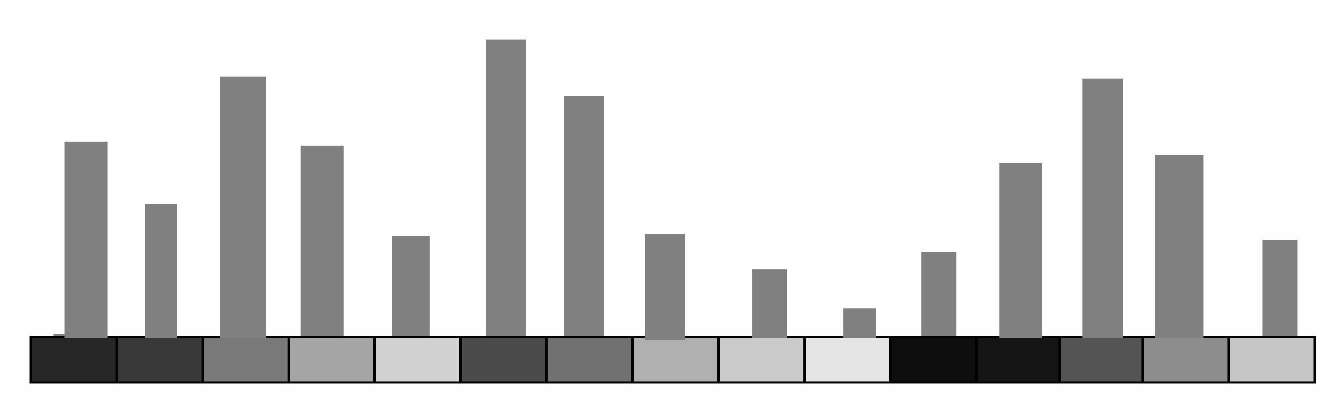
Query neighbourhood

Indexing images

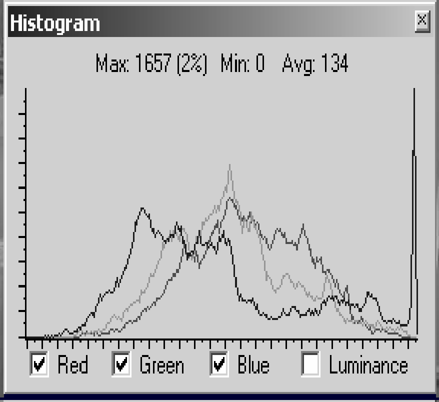
- **Colour spaces**
  - The most common and intuitive colour space is the RGB (Red Green Blue) colour space
    - ➔ **Every perceivable colour can be obtained as the sum of three degree of RGB**

Image indexing

- **Colour histograms**
  - The colour spectrum is divided into *n* bins
  - The value contained in each bin is proportional to the amount of pixels having the colour of that bin



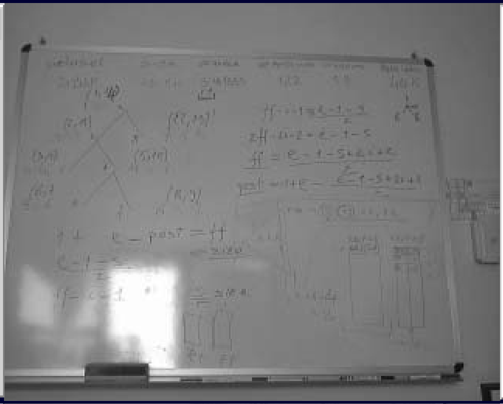
Indexing images

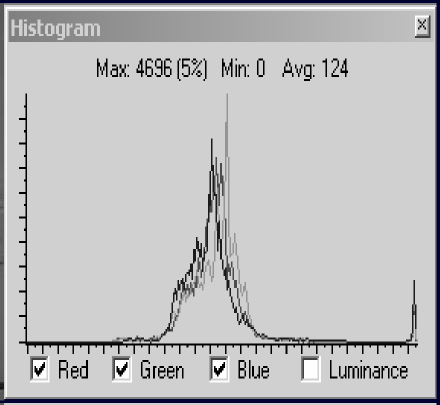


Histogram

Max: 1657 (2%) Min: 0 Avg: 134

☑ Red ☑ Green ☑ Blue □ Luminance

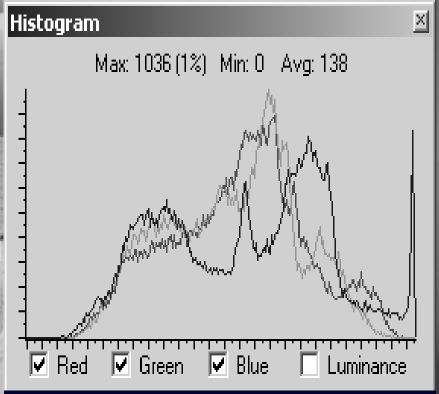




Histogram

Max: 4696 (5%) Min: 0 Avg: 124

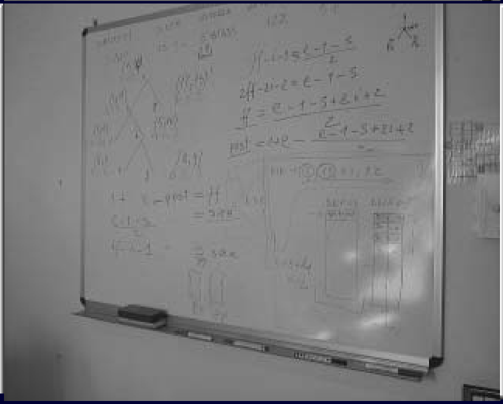
☑ Red ☑ Green ☑ Blue □ Luminance

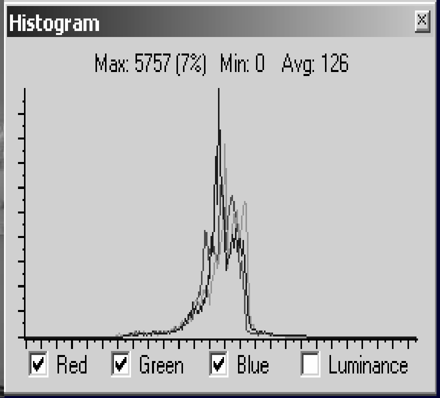


Histogram

Max: 1036 (1%) Min: 0 Avg: 138

☑ Red ☑ Green ☑ Blue □ Luminance





Histogram

Max: 5757 (7%) Min: 0 Avg: 126

☑ Red ☑ Green ☑ Blue □ Luminance

Indexing images

- **Problems with RGB:**
  - Colours that are close in the RGB colour space can be distant for the human perception

Indexing images

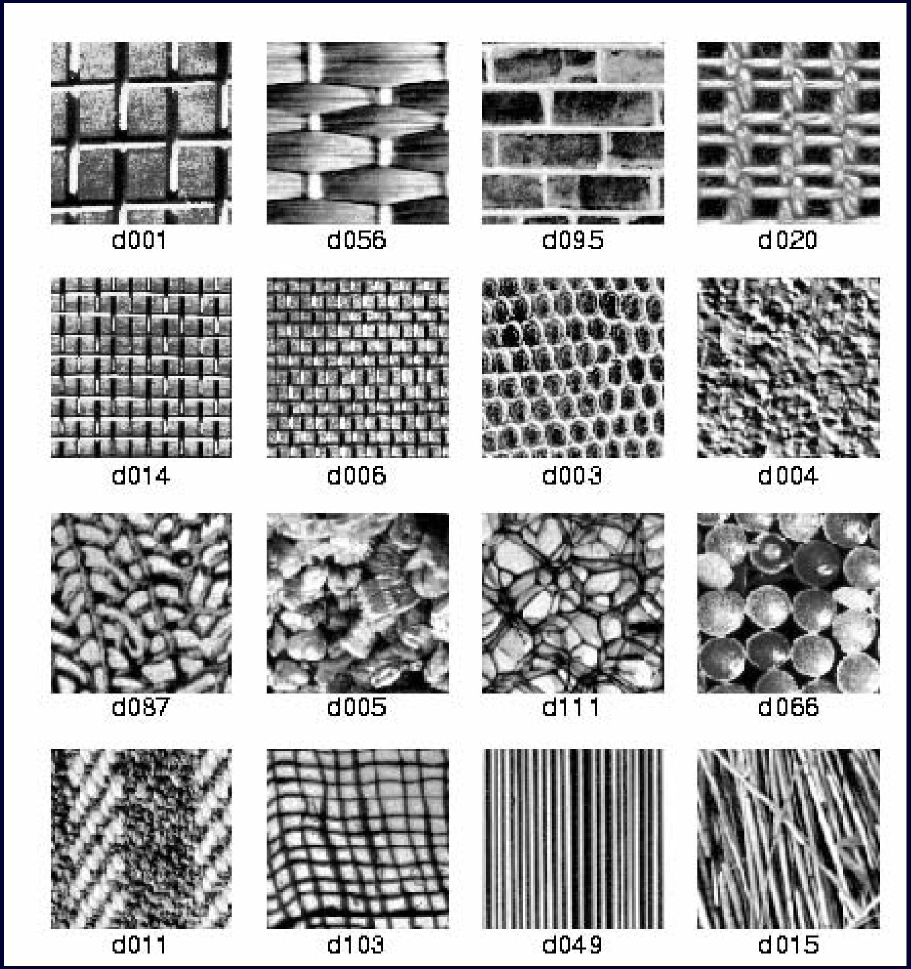
- **Wanted properties of colour spaces:**
  - Uniformity
    - → **Close colours are also perceived as similar**
  - Completeness
    - → **All perceivable colours are representable**
  - Compactness
    - → **No redundancy**

Indexing images

- **Other colour spaces:**
  - HSV
    - ➔ **Hue:Tint of the colour**
    - ➔ **Saturation:Quantity of colour**
    - ➔ **Value (Brightness):Quantity of light**
  - YIQ, YUV, YCrCb, etc.

Indexing images

- **Textures:**



Indexing images

- **Textures:**
  - Homogeneous patterns
  - Spatial arrangement of pixels

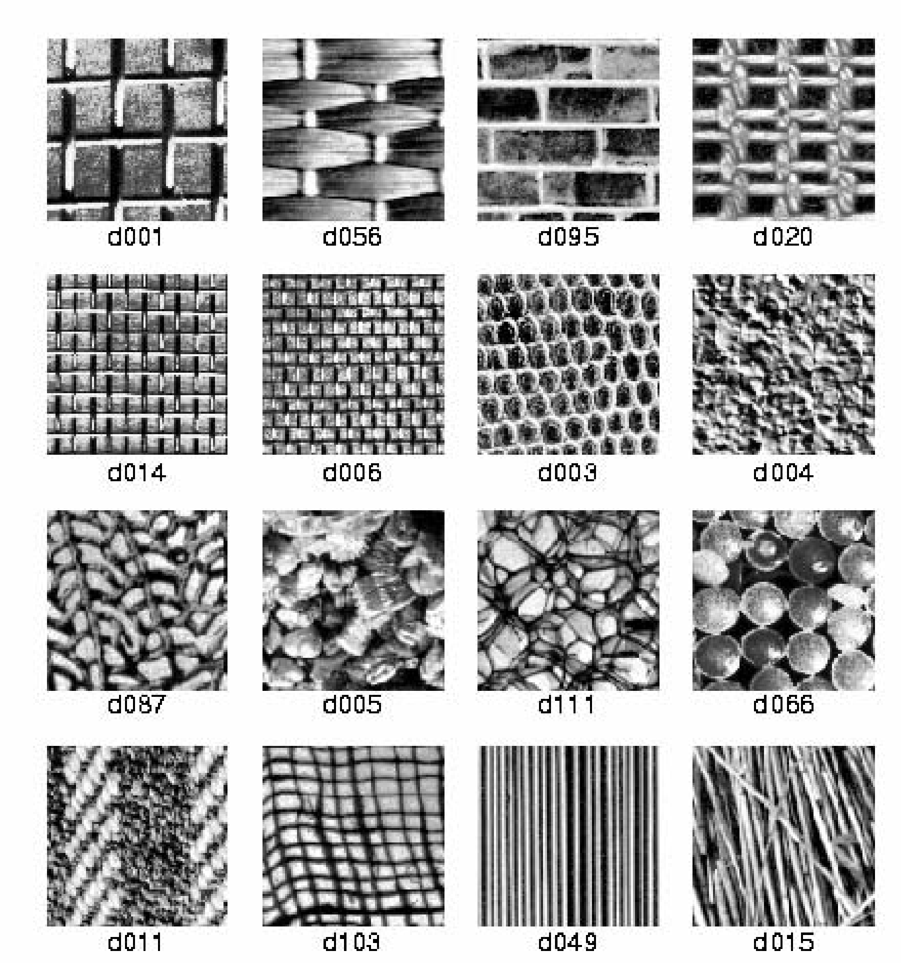
    ➔**Colour is not enough to describe**

Indexing images

- **Textures descriptions are obtained by using statistical methods**
  - Spatial distribution of image intensity
  - Several methods exists
  - Texture descriptions can also be represented as histograms (vectors)

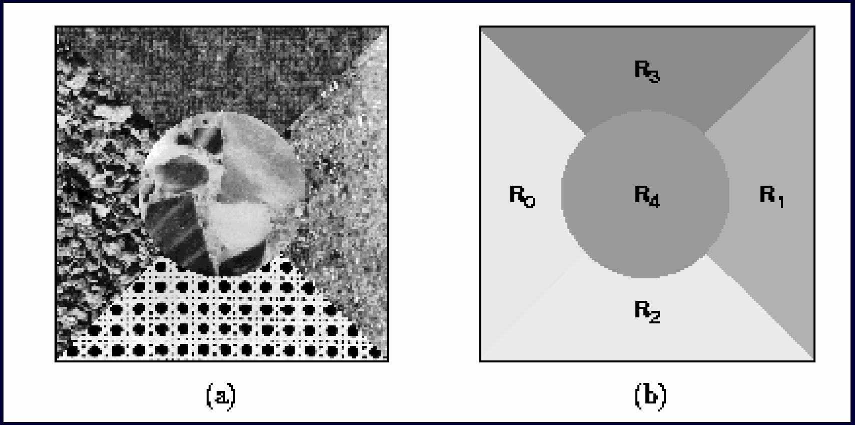
Indexing images

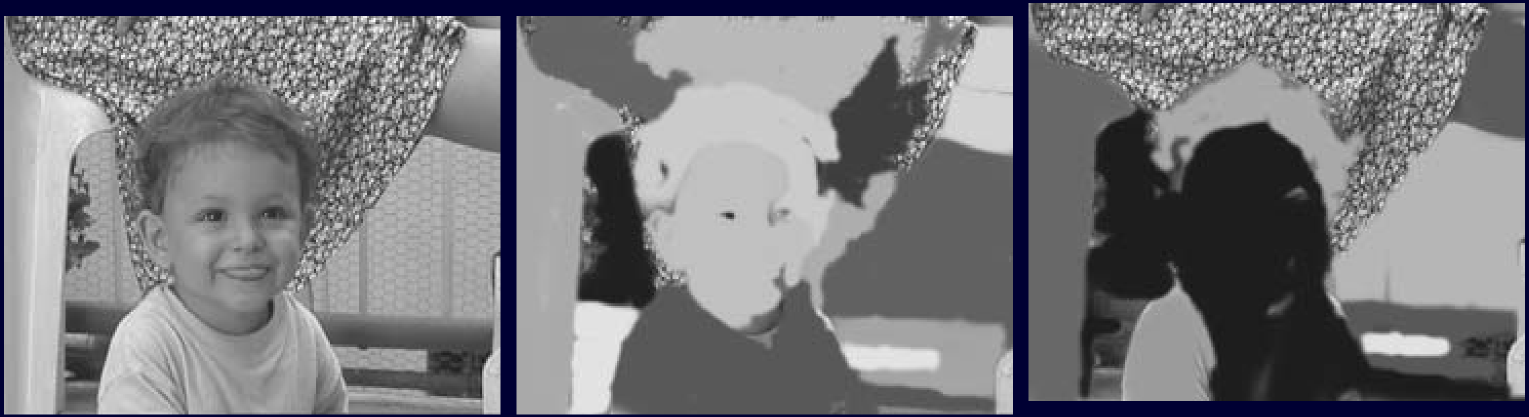
- **Widely used features for textures are the Tamura features:**
  - Contrast
    - **➔ Distribution of pixel intensity**
  - Coarseness
    - **➔ Granularity of a texture**
  - Directionality
    - **➔ Dominant direction of the texture**



Indexing images

- **Shapes:**
  - Region extraction
  - Segmentation





Indexing images

- **Colour histograms and textures can be computed for individual regions in addition to entire images**
  - Global features
    - ➔ **Search for images**
  - Local features
    - ➔ **Search for regions in images**

- **Spatial relationships between regions give also additional information**
  - Search for images having specific characteristics