Multimedia retrieval

- sample tasks
  - find copies and recordings of music/video
  - find all images/videos containing a person
  - find audio/video containing a spoken keyword
  - find music in a similar style
  - find all books/newspaper articles mentioning a keyword
- sample applications
  - personal media management (duplicate removal, snapshots, …)
  - Forensic applications (illegal copies, pornography, evidence, …)
  - social science research (finding people, objects, ...)

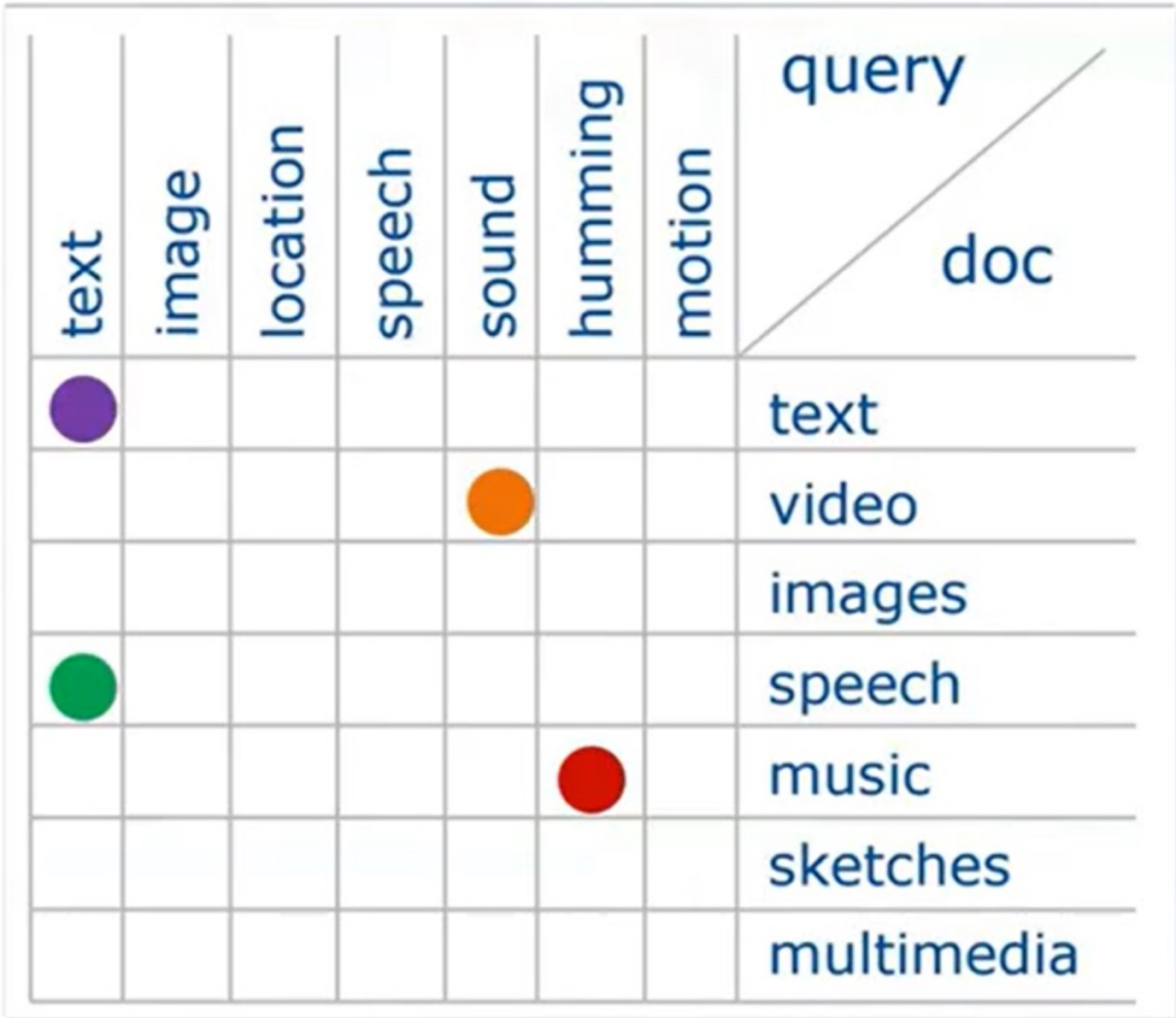
What is Multimedia?

Multimedia Queries

Web based image search

Nirma University auditorium

New search types



Conventional text retrieval

Type “terror attacks” and get all CCNIBN news

Hum a music piece and get all similar music

Roar like a Lion and get a wild life documentary

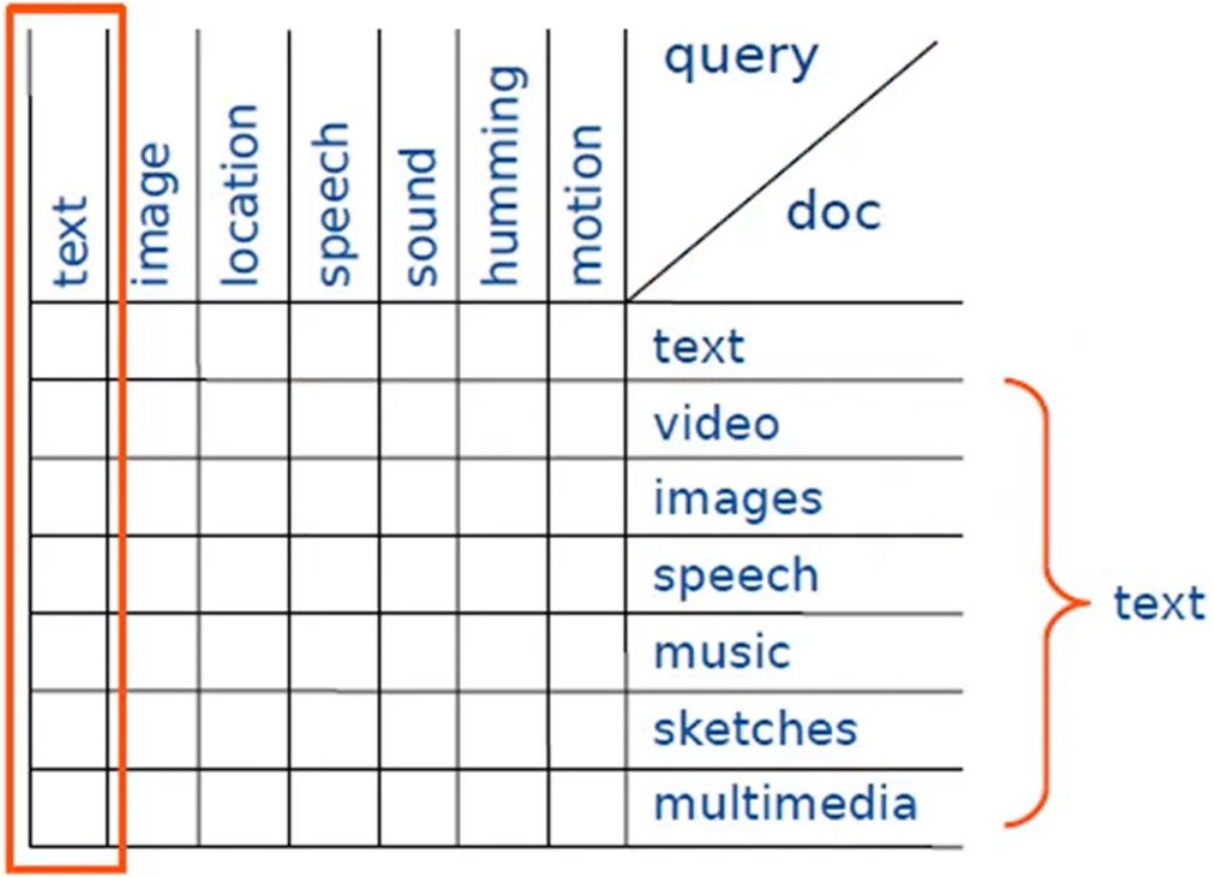
Challenges in MMIR

- Semantic Gap
- Ploysemy
- Fusion Problem: how to combine possibly conflicting evidence of two images similarity?
- Responsiveness: Naïve comparison of query feature vectors to the database feature vectors requires linear scan through the database.

Polysemy

- Metadata are pieces of information about a multimedia object that are not strictly necessary for working with it, but that are useful to
  - describe resources so they can be indexed, classified, located, browsed and found
  - store technical information, such as data formats and compression schemes
  - manage resources such as their rights or where they are currently located
  - record preservation actions
  - create usage trails, eg, which section of a video has been watched how many times

Piggyback retrieval



Content based image retrieval

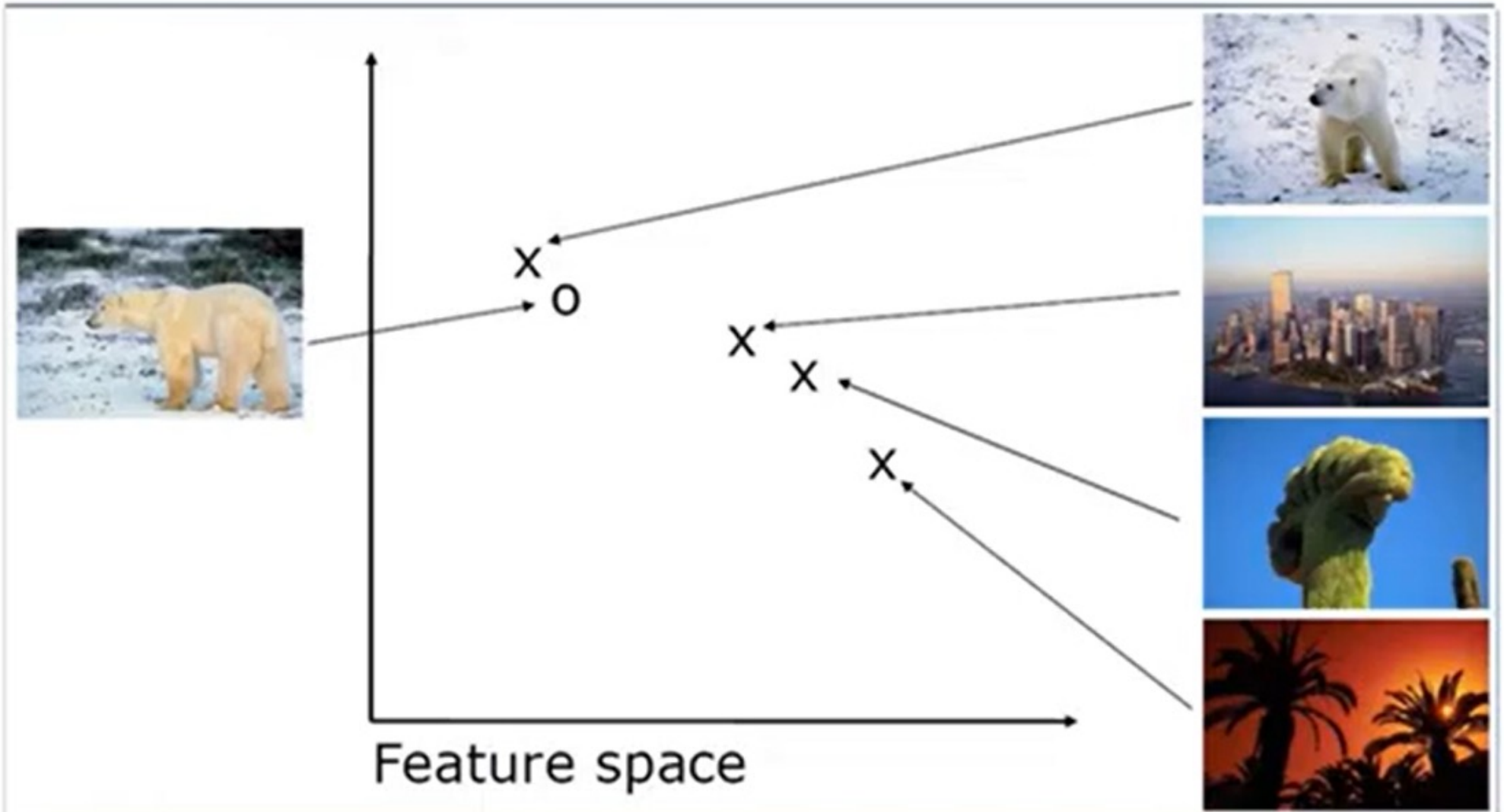
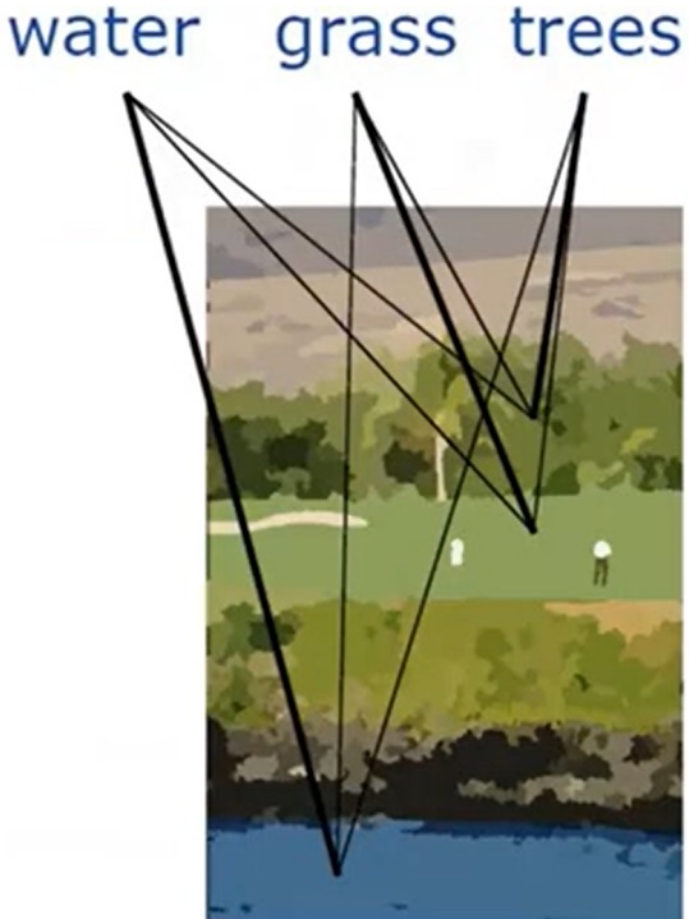


Image Annotation



Automatic Annotation (ML)

- Probabilistic models:
- maximum entropy models
- models for joint and conditional probabilities
- evidence combination with Support Vector Machines

A simple bayesian classifier

$$P(w|I) = \frac{P(w,I)}{P(I)} = \frac{\sum_J P(w,I|J)P(J)}{\sum_J P(I|J)P(J)}$$

$$= \frac{\sum_J P(I|w,J)P(w|J)P(J)}{\sum_J \sum_w P(I|w,J)P(w|J)P(J)}$$

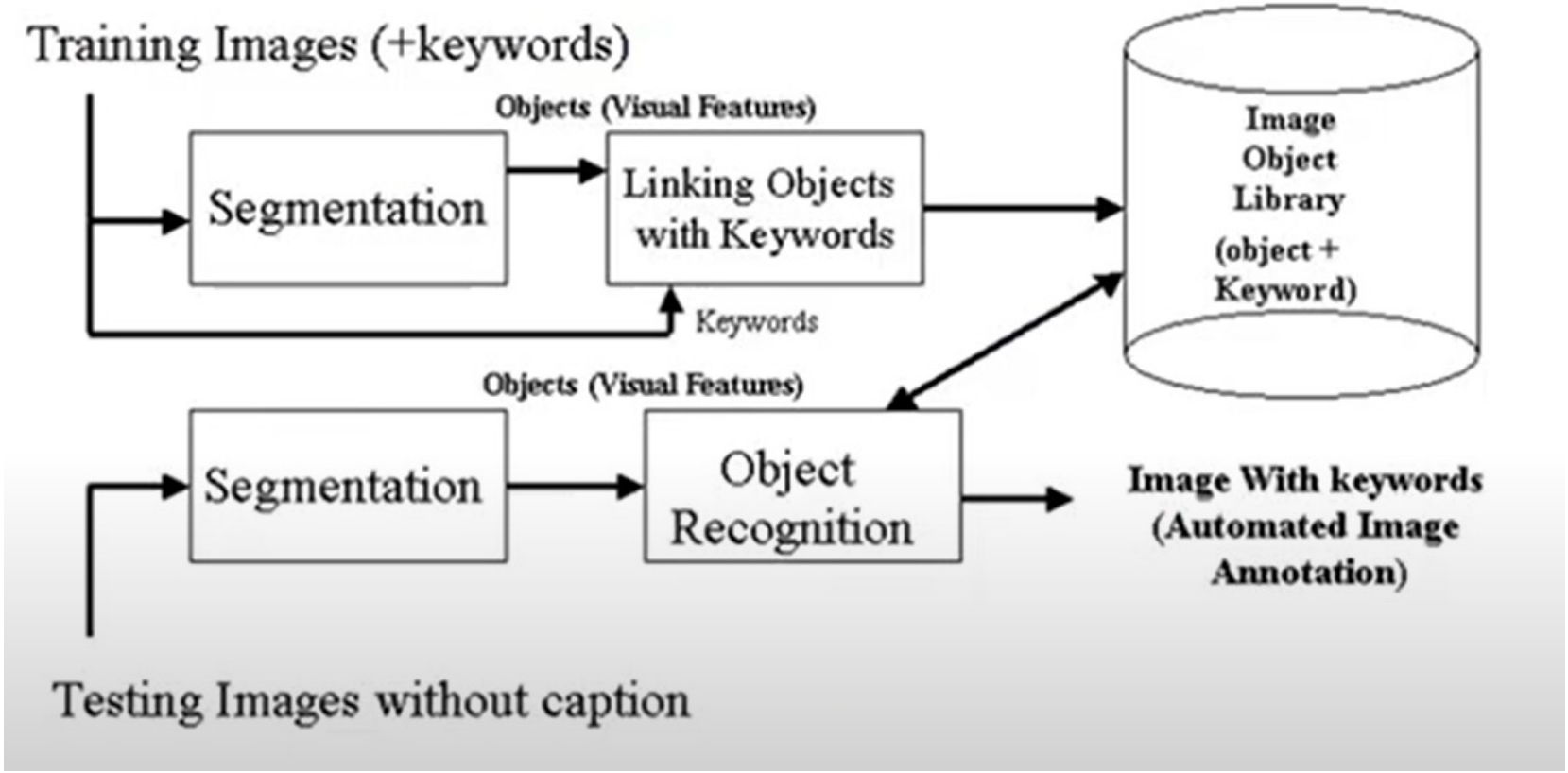
Use training data J and annotations w

P(w|I) is probability of word w given unseen image I

The model is an empirical distribution (w,J)

- Keyword is associated with the visual term
- Calculate the probabilities of translation of this visual term into water, into the grass and into the tree and whichever probability is highest that is the correct annotation of this visual term.
- Training data is required. The initial set of annotations is required.

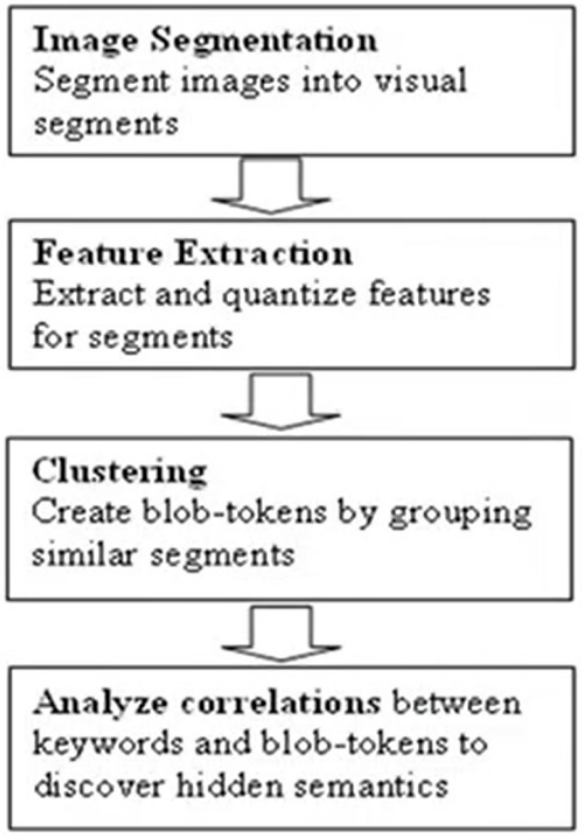
Annotation


Training Images (+keywords)
Objects (Visual Features)
Segmentation
Linking Objects with Keywords
Keywords
Image Object Library (object + Keyword)
Objects (Visual Features)
Segmentation
Object Recognition
Image With keywords (Automated Image Annotation)
Testing Images without caption

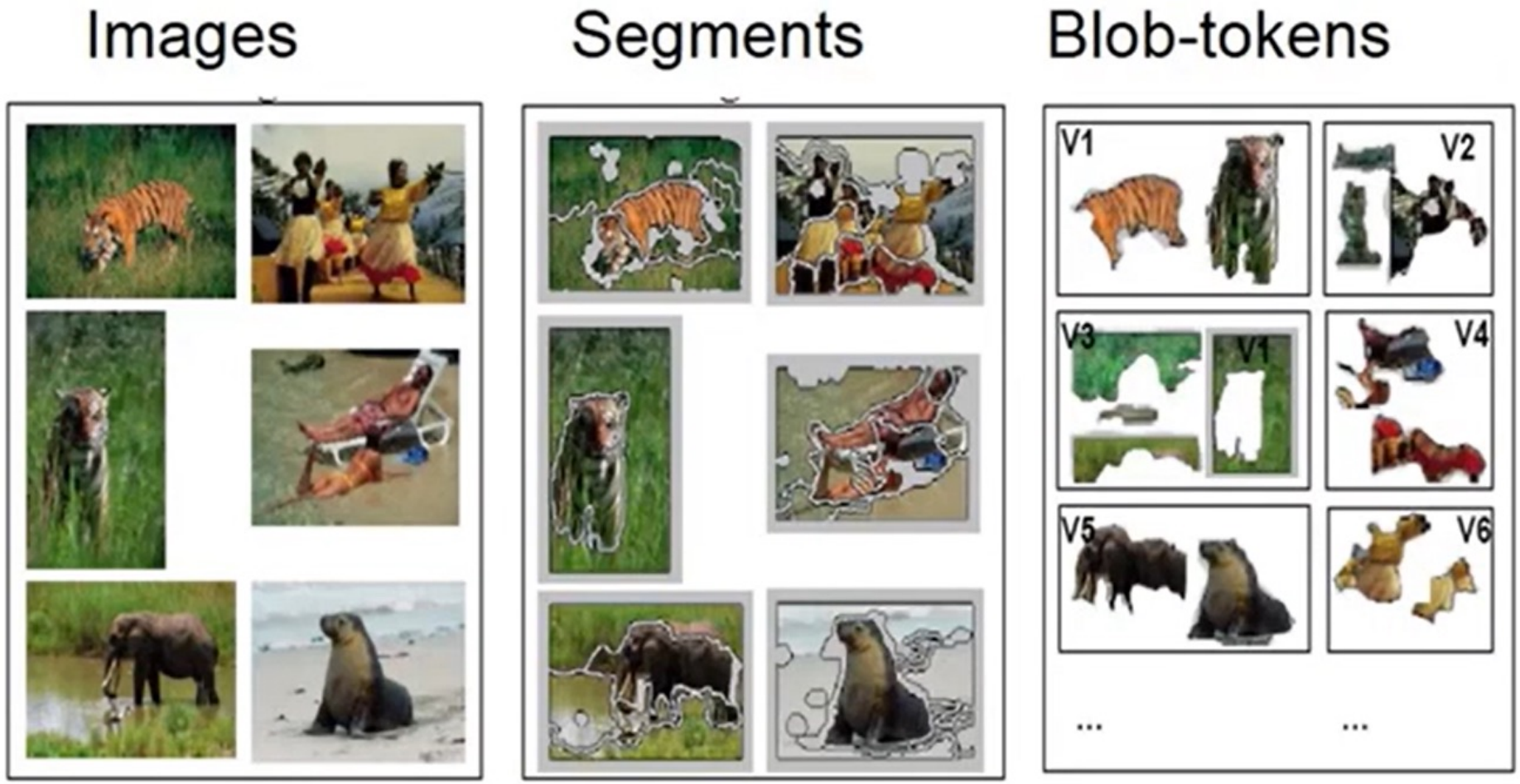
Annotation

Major steps:

- **Segmentation** into regions
- **Clustering** to construct **blob-tokens**
- Analyze **correspondence** between key words and blob-tokens
- **Auto Annotation**



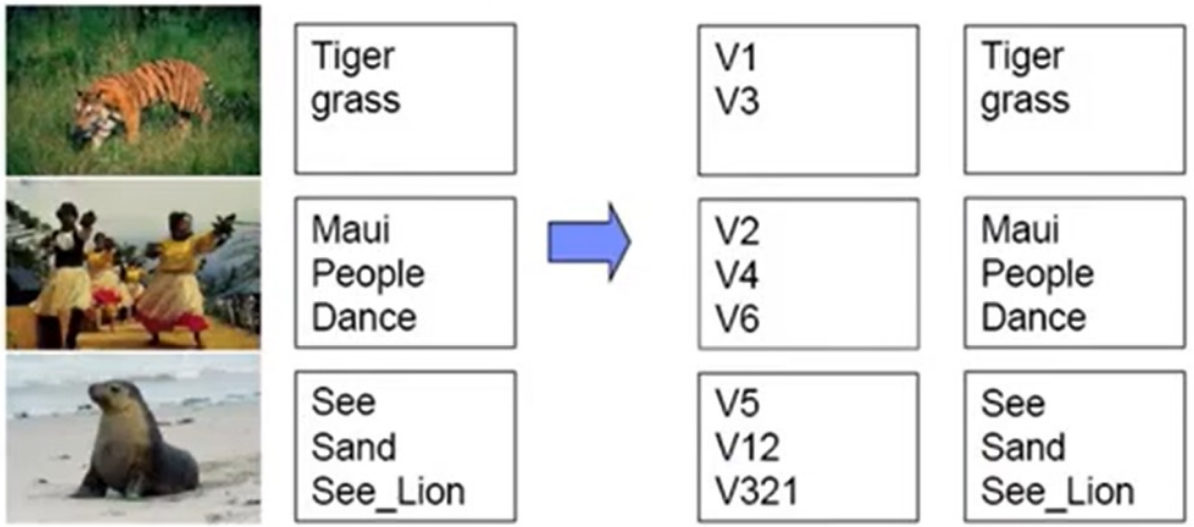
Annotation segmentation and Clustering



Annotation linking

Our purpose is to find correspondence between words and blob-tokens.

P(Tiger|V1), P(V2|grass)…



Feature Extraction and Clustering

- Feature Extraction:
  - Color
  - Texture
  - Shape
- K-means clustering: To generate finite visual terms. Each cluster's centroid represents a visual term.

Co-Occurrence Model

- Mori et al. 1999
- Create the co-occurrence table using a training set of annotated images
- Tend to annotate with high frequency words
- Context is ignored
  - Needs joint probability models

	w1	w2	w3	w4
V1	12	2	0	1
V2	32	40	13	32
V3	13	12	0	0
V4	65	43	12	0

$P(w1 \mid v1) = 12/(12+2+0+1)=0.8$

$P(v3 \mid w2) = 12/(2+40+12+43)=0.12$

Correspondance : Translation Model(TM)

"sun sea sky"

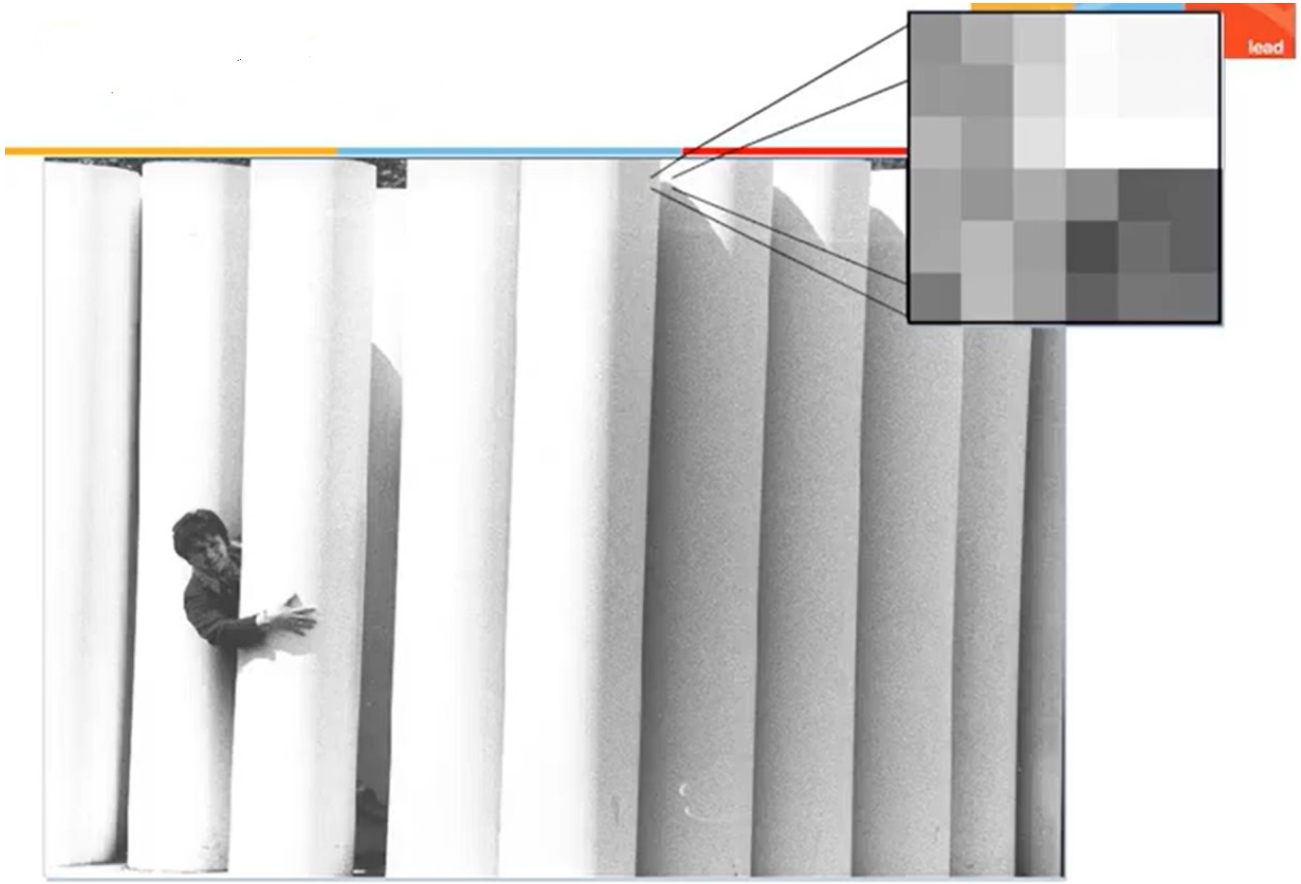
"sun sea sky"

$$\Pr(w|v) = \sum \Pr(w,a|v)$$

Image Description

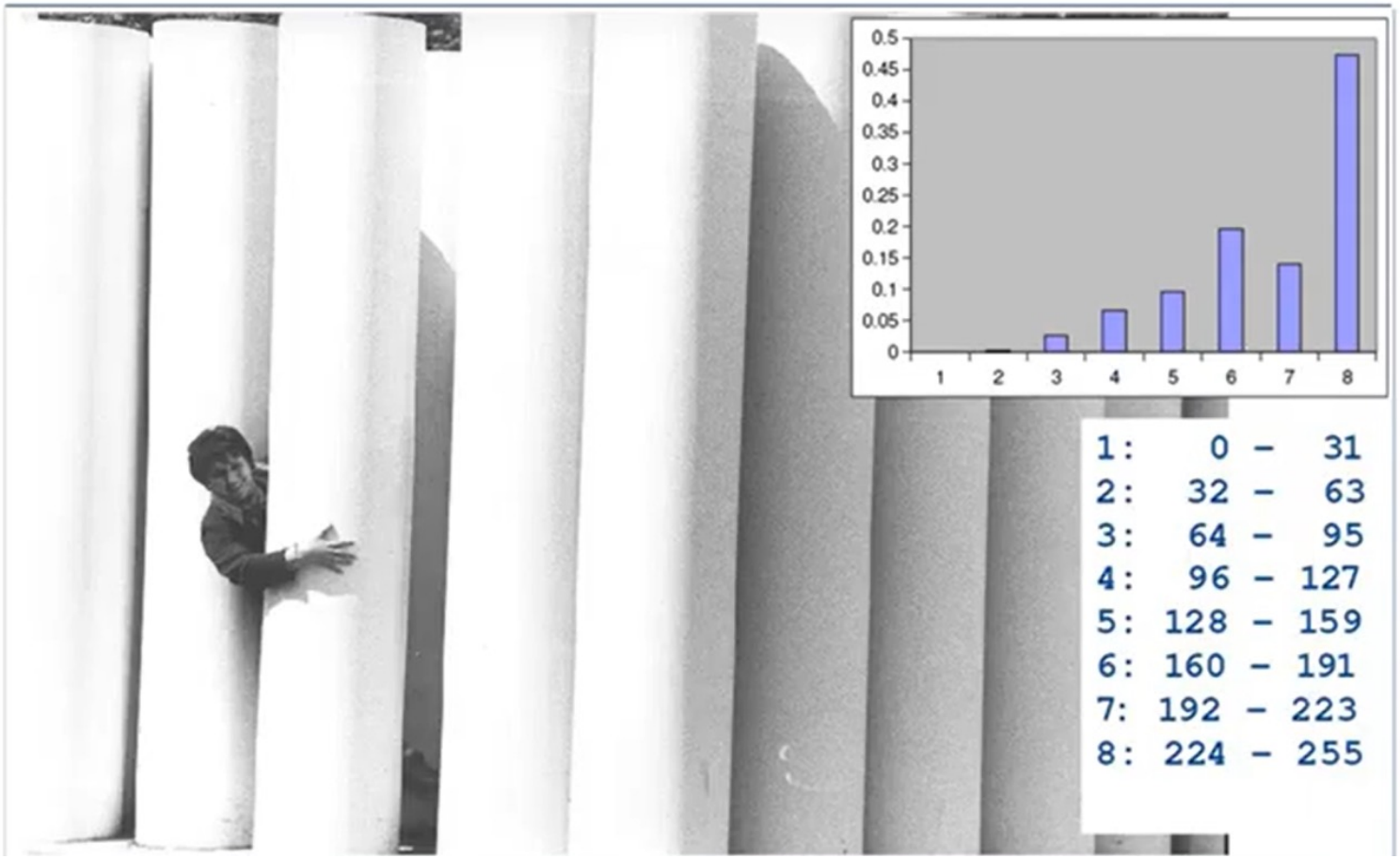
- Subjective interpretation of content: means different things to different people
- Different features for different applications
- Colour is important of out-door image but not for X-rays, CT, MRI etc.
- Motion features are sometimes important (ultrasound)
- Different systems for different applications

Digital Image

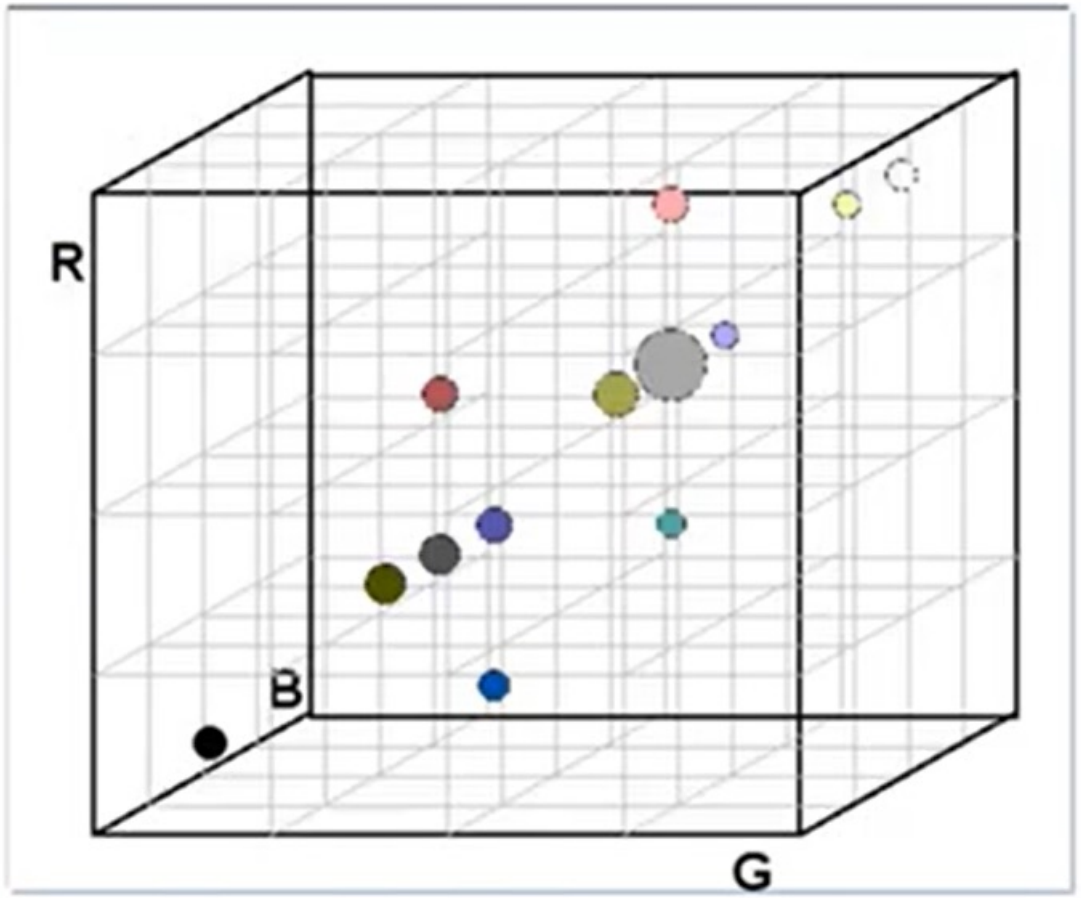


145	173	201	253	245	245
153	151	213	251	247	247
181	159	225	255	255	255
165	149	173	141	93	97
167	185	157	79	109	97
121	187	161	97	117	115

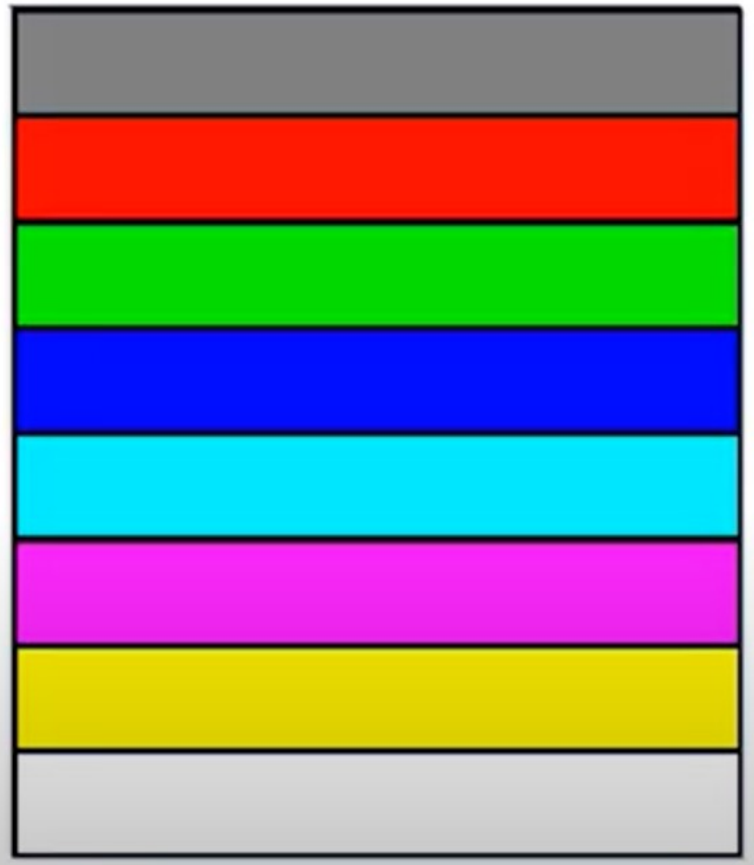
Histogram


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1 2 3 4 5 6 7 8
1: 0 - 31
2: 32 - 63
3: 64 - 95
4: 96 - 127
5: 128 - 159
6: 160 - 191
7: 192 - 223
8: 224 - 255

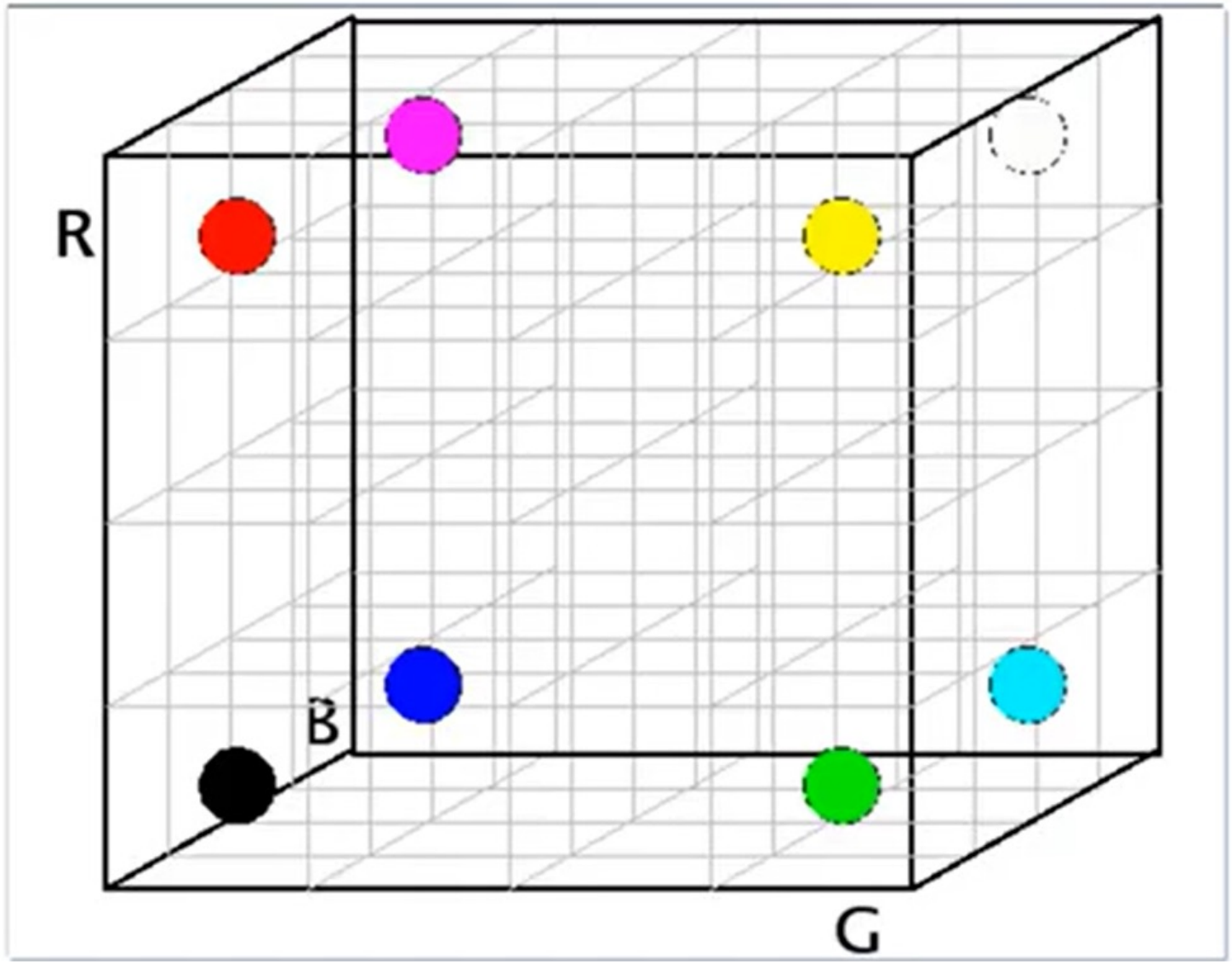
Color Histogram



Color code

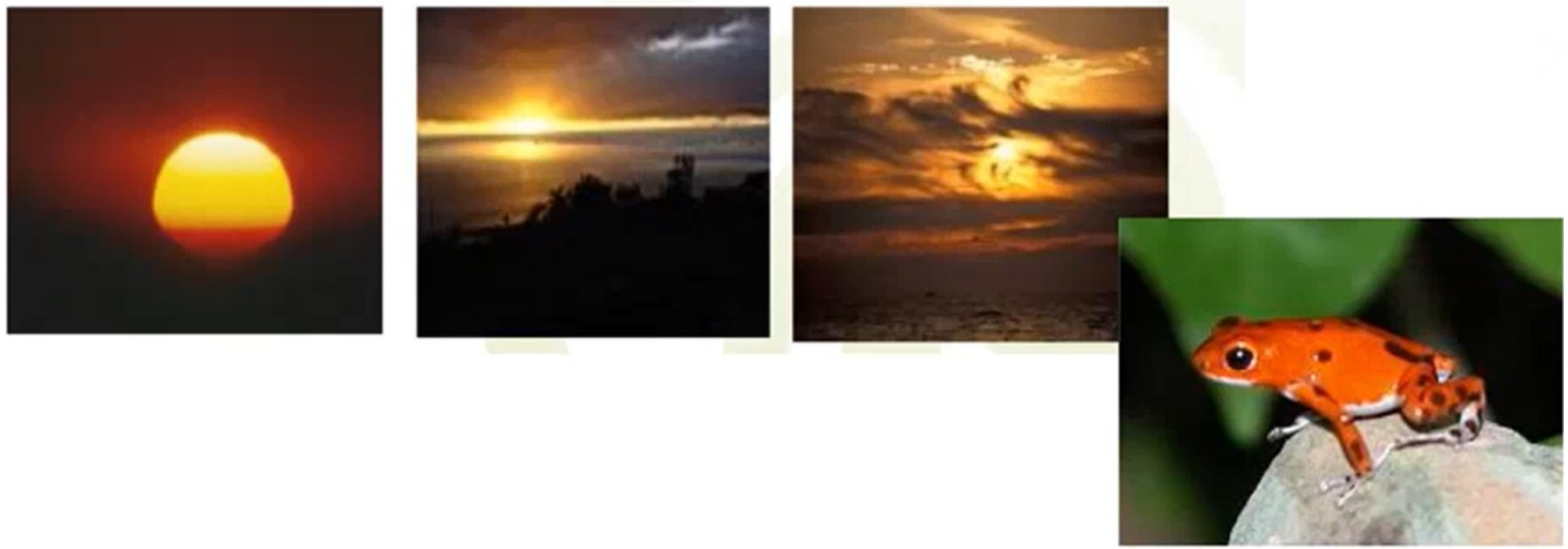


R	G	B	
0	0	0	black
255	0	0	red
0	255	0	green
0	0	255	blue
0	255	255	cyan
255	0	255	magenta
255	255	0	yellow
255	255	255	white


R
B
G

```
Off-line, for each image
    create histogram with a bin for each color
    initialize each bin counter = 0
        for each pixel in image:
            increment bin counter corresponding to pixel
            color
        end
```

- **Assumption:** If two images share similar colors then also their content may be similar
- Loss of information through low-level features
- Example: red sunset (orange, yellow)



Simple Histogram Distance Measure

- The distance between the histogram of the query image and images in the database are measured
- Image with a histogram distance smaller than a predefined threshold are retrieved from the database
- The simplest distance between images I and H is the L-1 metric distance as $D(I,H) = \text{sum } |I-H|$

Comparison of Image

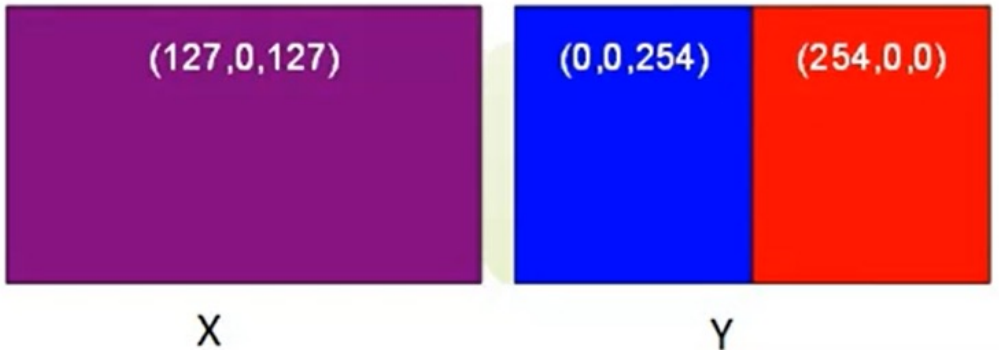
- Compare images based on the color? Extract color features first –
  - Each pixel of an image contains color information
- Images consist of many pixels
  - Pixel by pixel?
- Aggregation for comparisons?
  - Average color
  - Color histograms
  - Color layout (regions)

Average Color method

- Comparison of 2 images x and y by using the Euclidean distance for the average color

$$d^2_{avg}(x,y) = \left(R_{avg}{}^x - R_{avg}{}^y\right)^2 + \left(G_{avg}{}^x - G_{avg}{}^y\right)^2 + \left(B_{avg}{}^x - B_{avg}{}^y\right)^2$$

- Very bad similarity measure
- E.g., magenta image and red blue image are the same according to average color



Problem with Average color method

- Perceptionally somewhat questionable ...
- But...
  - Quick and easy to calculate and compare
- Best to use as a filter: exclude images
  - Dominant color influences the average color, the opposite is not valid
  - E.g., search for mostly blue images: exclude all images with red, yellow or green color averages

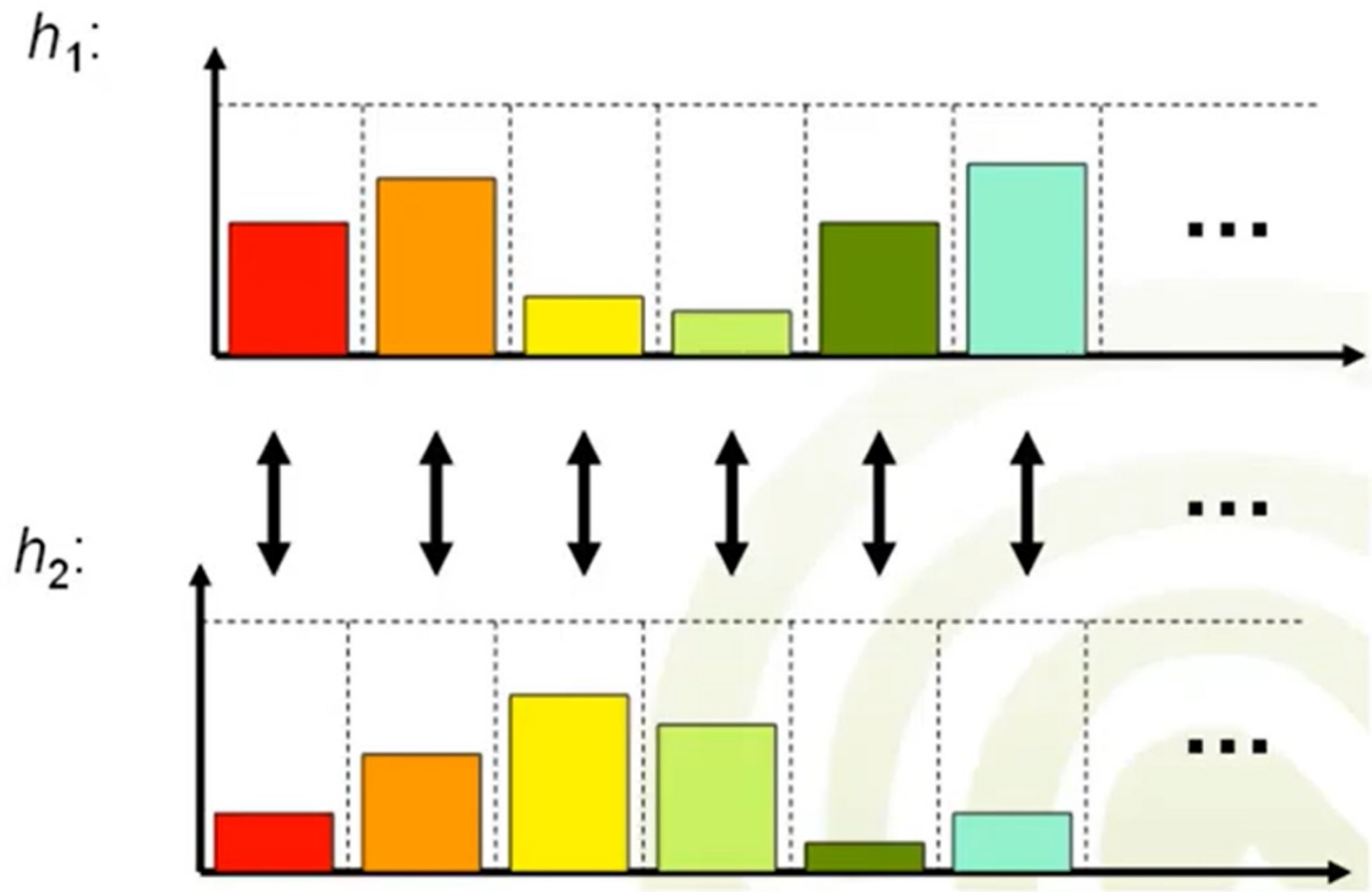
Histogram comparison method

- Given: histograms h_1 and h_2
- **Minkowski distance** with parameter r:

$$\mathrm{d}_r(h_1, h_2) := \sum_{i \in C} \left| h_1(i) - h_2(i) \right|^r$$

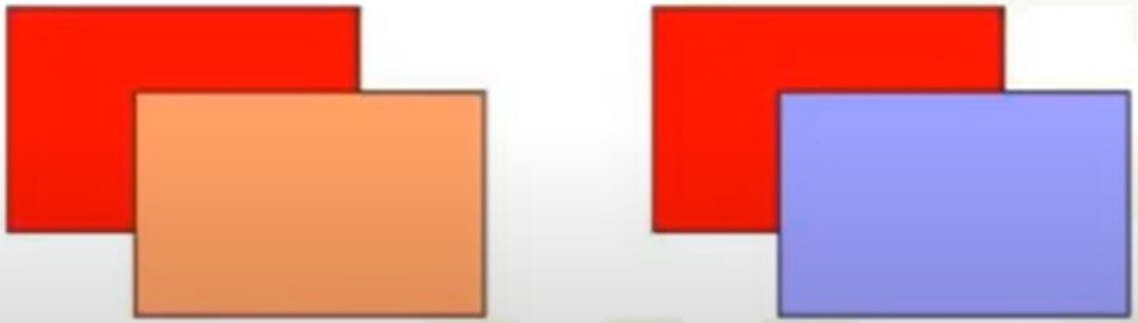
- $r = 1$: Histogram-L $_1$-norm
  (also: city block distance, Manhattan distance)
- $r = 2$: Histogram-L $_2$-norm (Euclidean)

Minkowski Distance



Problem with Minkowski Distance

- It is efficient to compute, but does not take the similarity of colors into account
  - The distance between a red and a bright red image is the same as between a red and blue one



- Works poorly in the case of **color shifts** because all columns are individually compared

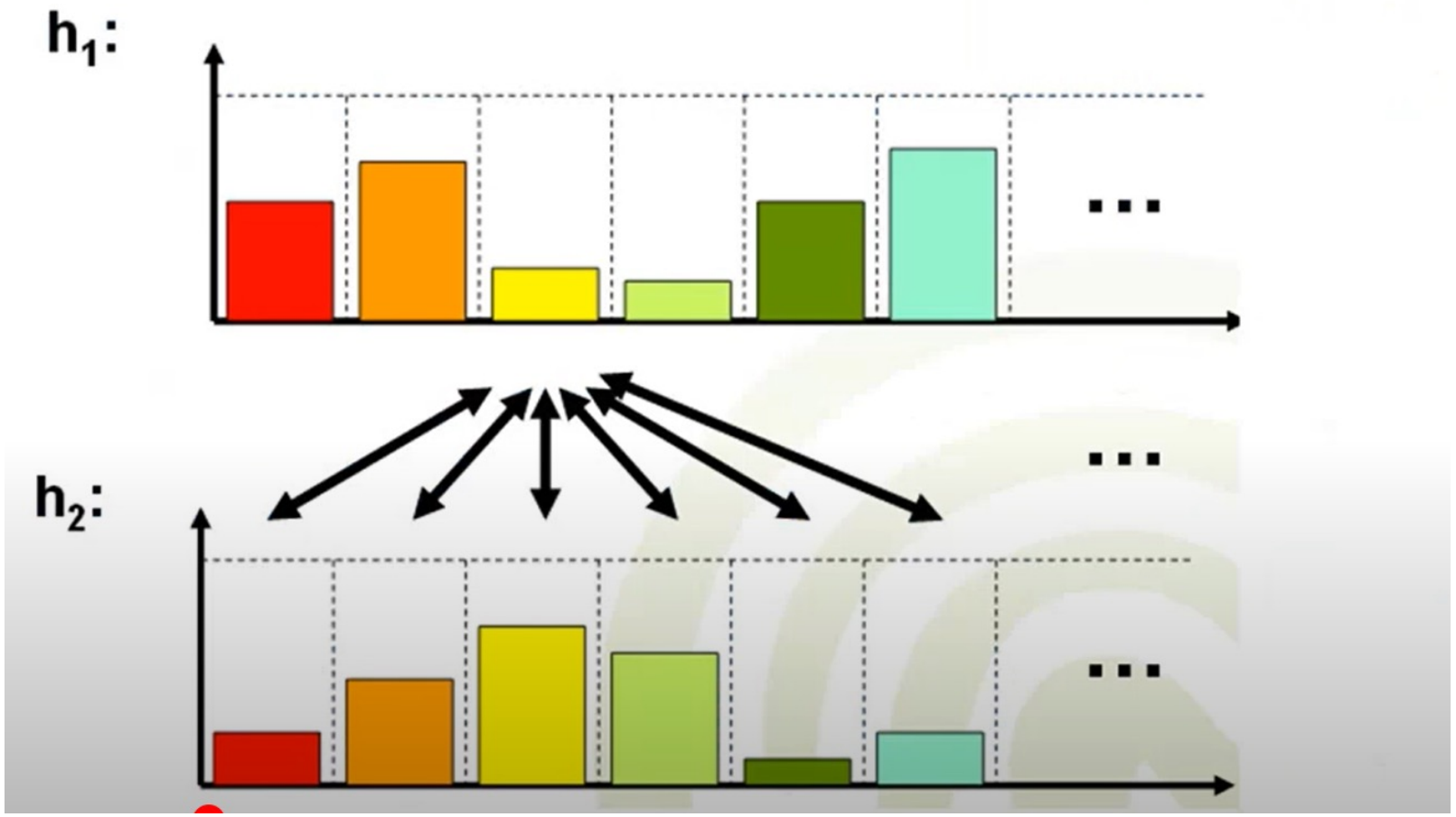
Quadratic distance measure(r=2)

- **Quadratic** distance measures
  - Evaluates the relationship between different colors in the histogram
  - **Cross-talk matrix:** A expresses pairwise similarity $a_{i,j}$ between color i and color j ($a_{i,i}=1$ and $a_{i,j}=a_{j,i}$):

$$d_A(h_1, h_2) = (h_1 - h_2)^T \cdot A \cdot (h_1 - h_2)$$

$$= \sum_{i \in C}^{k} \sum_{j \in C}^{k} a_{i,j} \cdot \big(h_1(i) - h_2(i)\big) \cdot \big(h_1(j) - h_2(j)\big)$$

Quadratic distance measure



Fusion of features spaces and query results

- Single Query Example with Multiple Features
  - Assume that there is a single query example q and that each multimedia document m gives rise to a number of low-level features f1(m), f2(m), . . ., fk(m), each of which would typically be a vector describing some aspect such as colour, texture, shape, timbre etc
- Combined Overall Distance
  - Most systems accumulate the distances of these features to the corresponding features of the query q in order to define an overall distance

$$D_w(m,q) = \sum_{i=1}^{r} w_i d_i(f_i(m), f_i(q))$$

  - between multimedia documents m and a query q. Here di(·, ·) is a specific distance function between the vectors from the feature i, and wi ∈ R is a weight for the importance of this feature

Take home message

- Since the features for multimedia objects are extracted using different tools it is nessary to under the underlying representation of these features.
- If necessary they have to be standardized.
- Efficient data structures have be used for indexing which can serious affect the performance of the system.
- Sometimes it may be essential to have one or many query objects as input at the same time which has be handled using different techniques explained.