CM3106 Multimedia

Content-based Retrieval

Dr Kirill Sidorov
SidorovK@cardiff.ac.uk
www.facebook.com/kirill.sidorov

Prof David Marshall
MarshallAD@cardiff.ac.uk

School of Computer Science and Informatics
Cardiff University, UK

Motivation

Suppose we want to search a multimedia database.

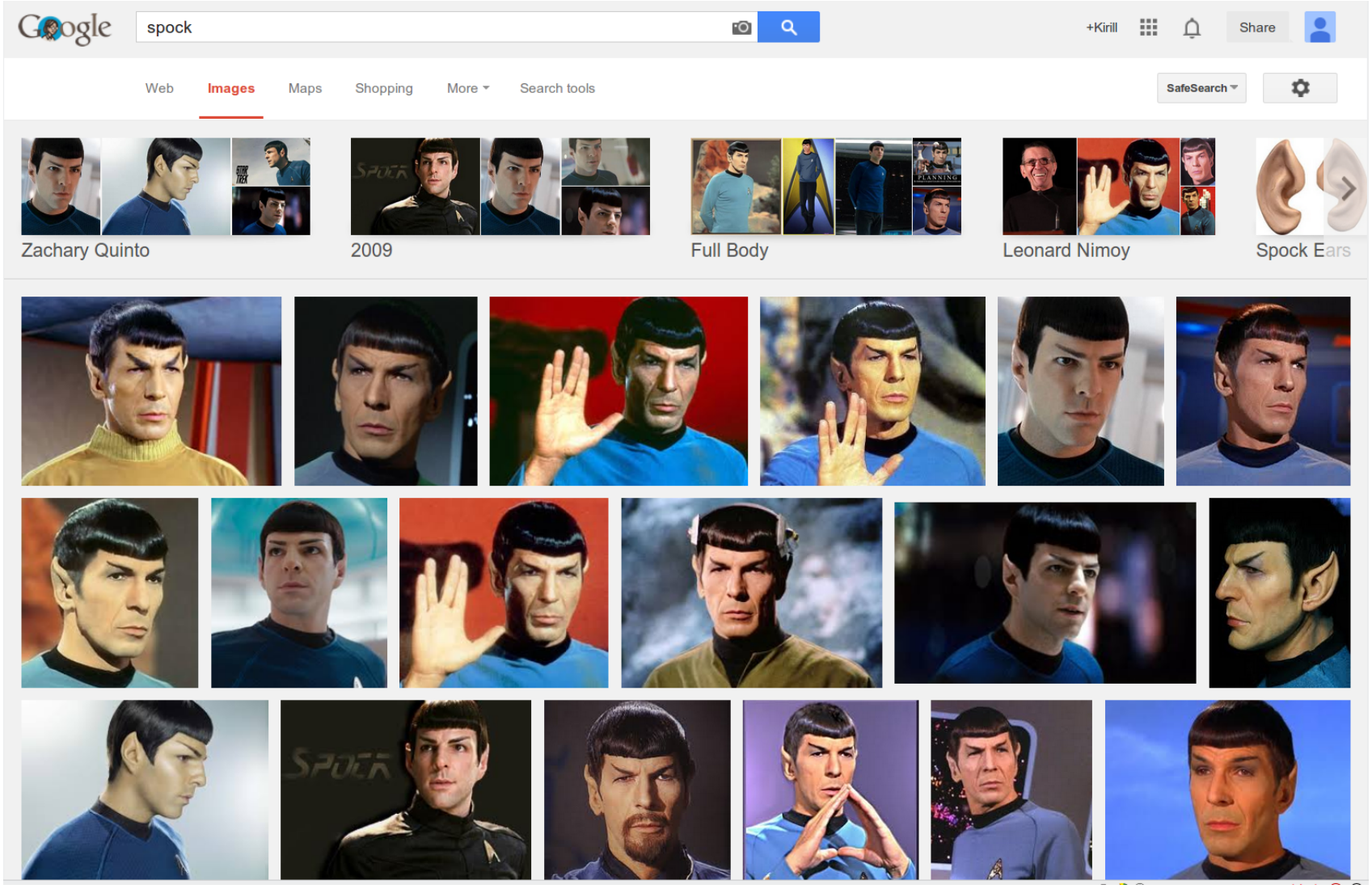
Applications:

- Medicine: find similar diagnostic images.
- Crime: find person according to mugshot, fingerprints, sketch, or verbal description.
- Art: search museum collection of paintings.
- Copyright: who used my images without permission?
- Retail: find shoes similar to these ones, only red.

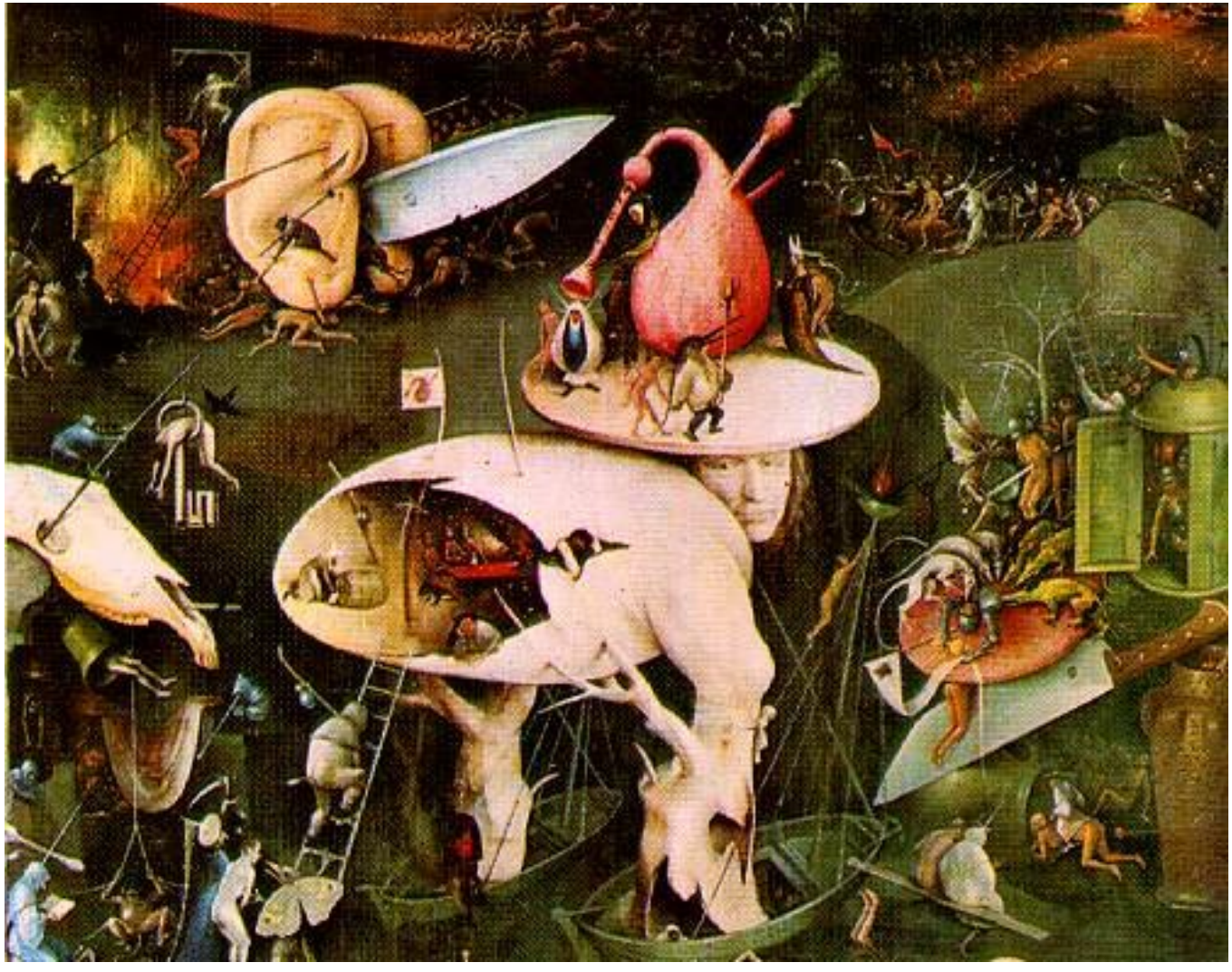
Traditional techniques

- Text-based multimedia search and retrieval:
  - Annotations (metadata).
  - File names. Keywords. Captions. Surrounding text. Photography conditions. Geo tags. Creation date.
  - Verbal portrait in the police database.
- Usually does a very good job provided the annotations are accurate and detailed.
- E.g. google image search, youtube video search.
- Disadvantages:
  - Manual annotation requires vast amount of labour.
  - Different people may perceive the contents of images differently: no objectivity in keywords/annotations.

Traditional techniques

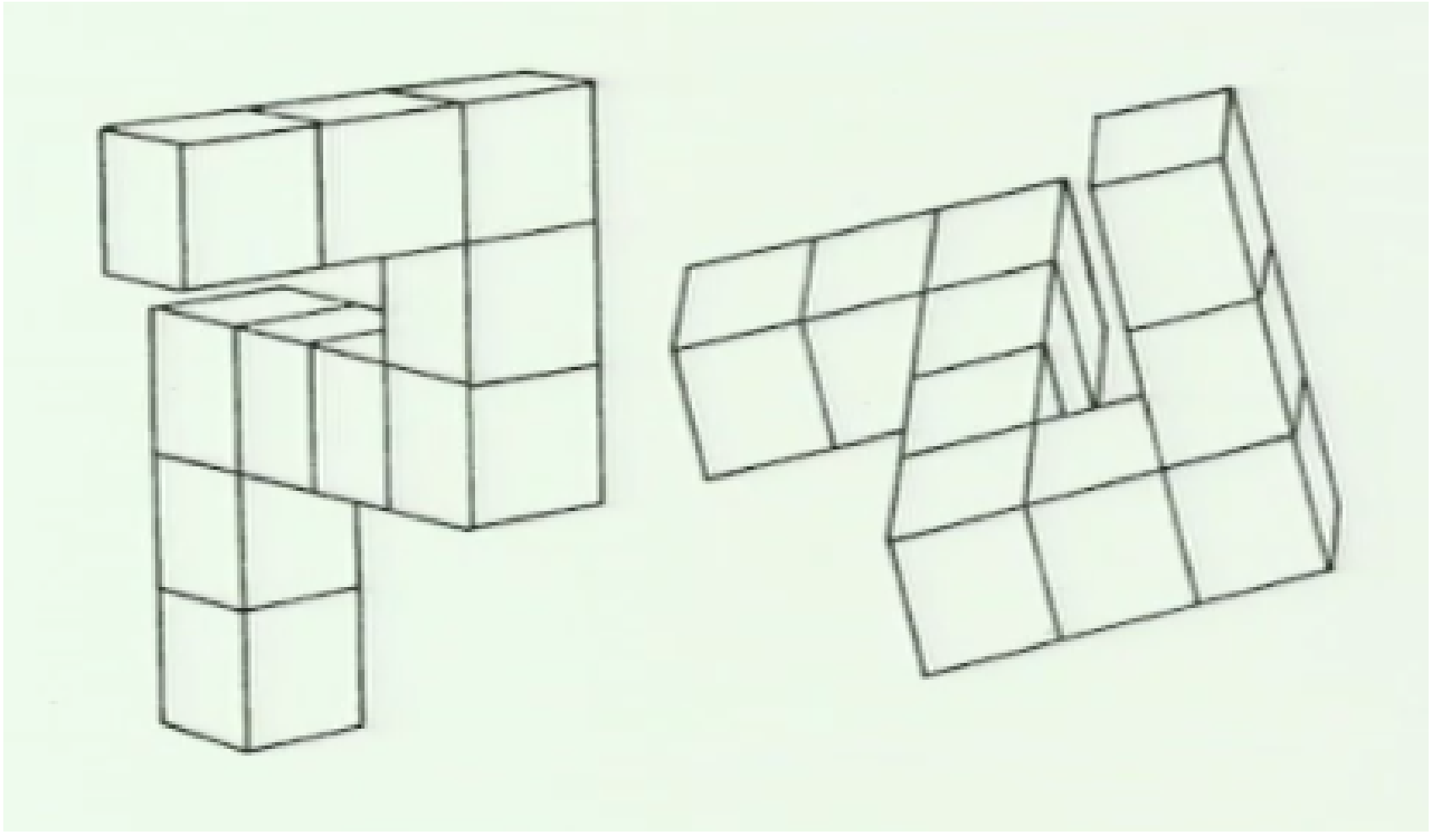


Traditional techniques

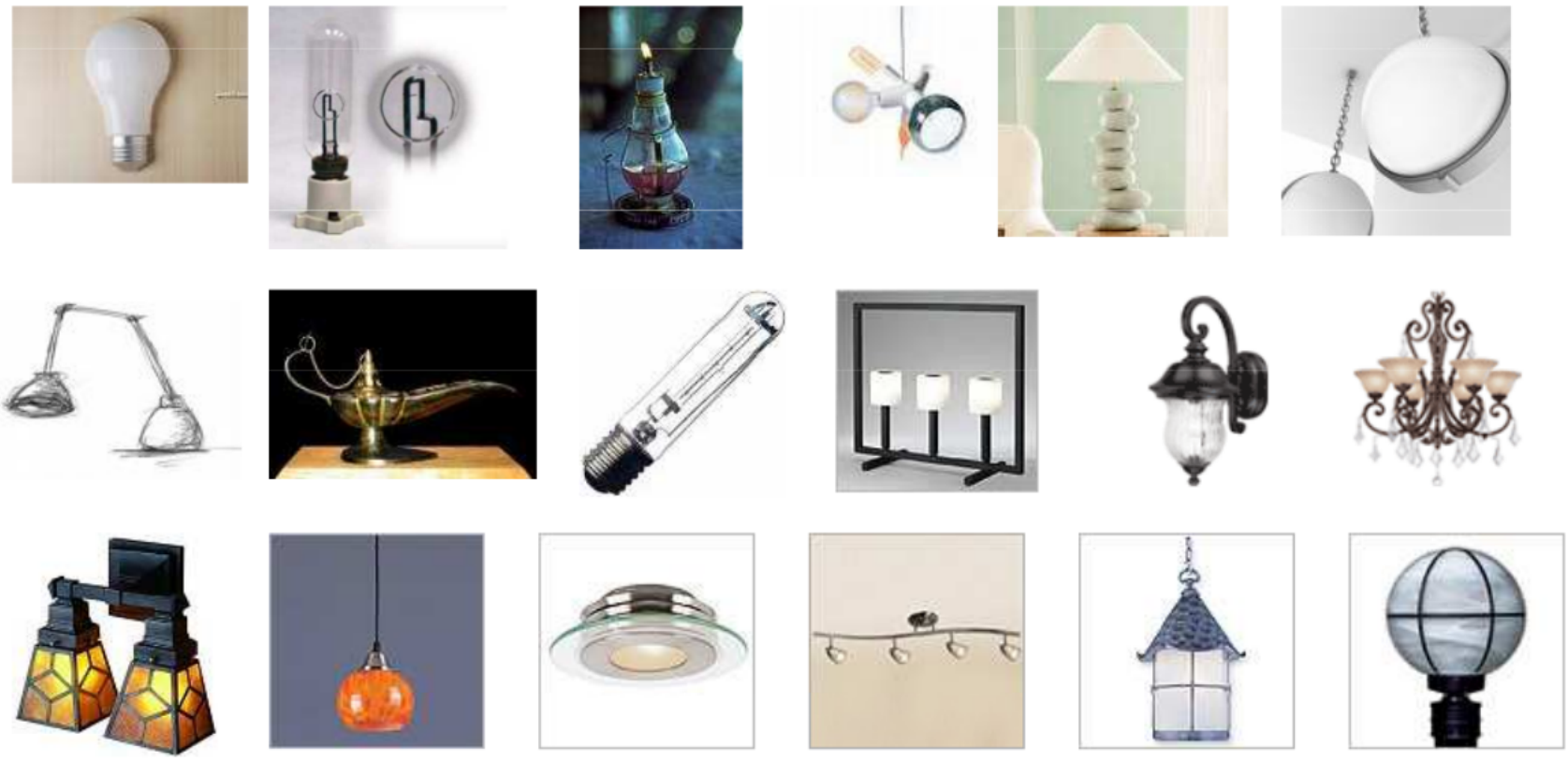


Describe in words what is happening in this image!

How do humans compare images?



How do humans compare images?



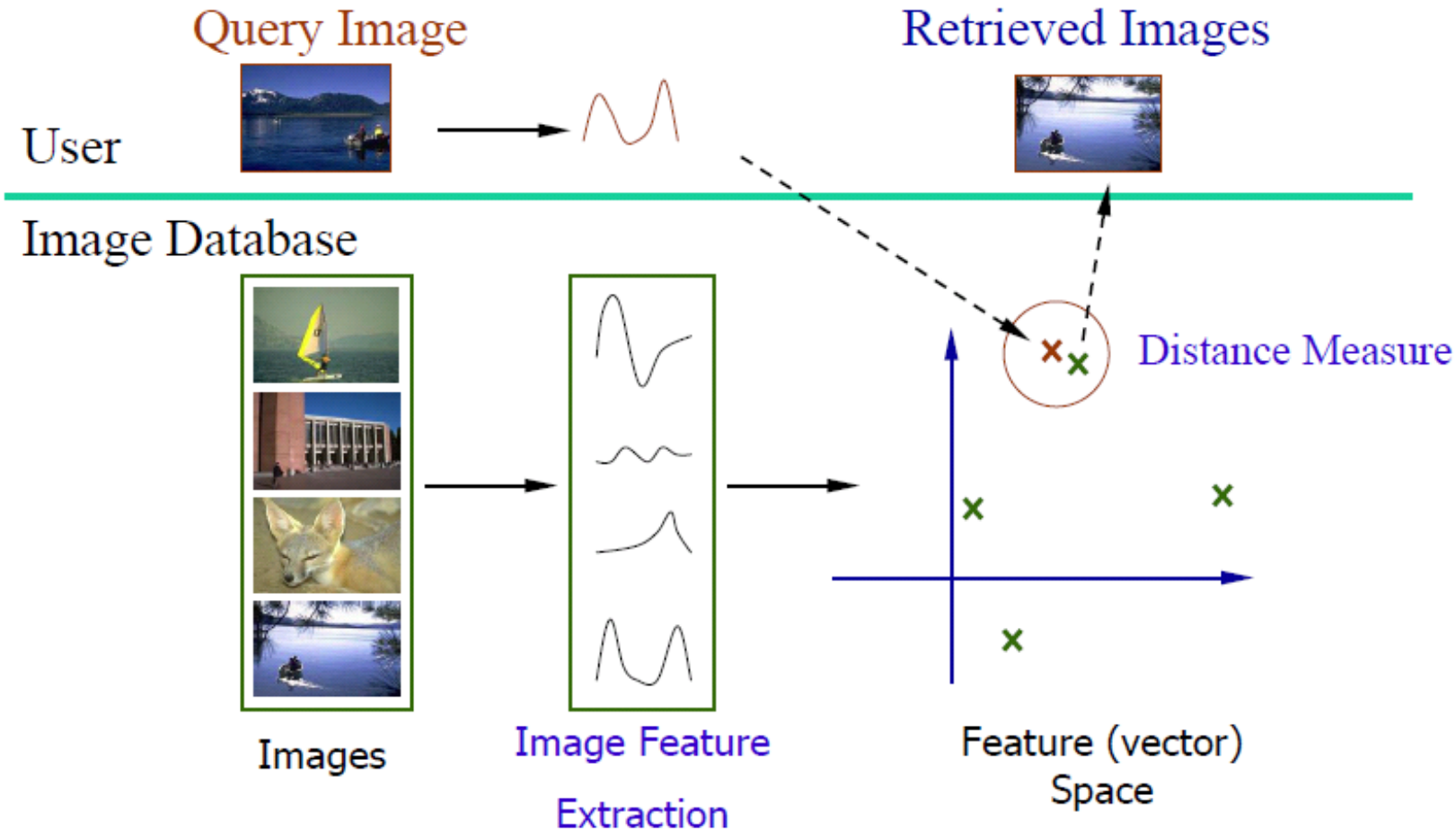
How do humans compare images?

How do humans compare images?

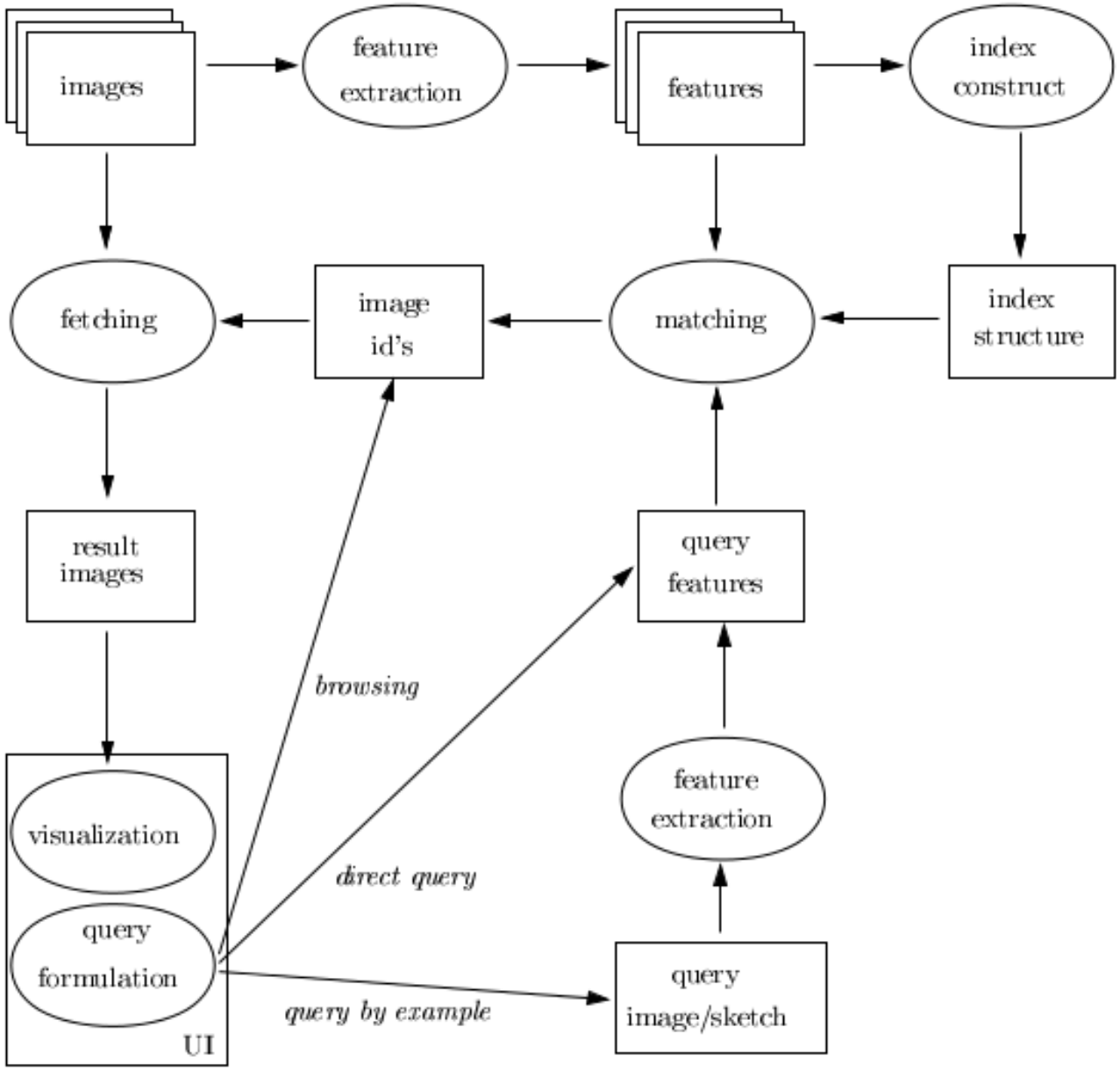
Content-based image retrieval

- Low-level: based on color, texture, shape features.
  - Find all images similar to given query image.
  - Search by sketch.
  - Search by features e.g. “find all green images with texture of leaves”.
  - Check whether image is used without permissions.
  - Images are compared based on low-level features, no semantic analysis involved.
  - A lot of research since 1990’s. Feasible task.
- Mid-level: semantics come into play
  - E.g. “find images of tigers”.
  - Very active and challenging research area.
- High-level:
  - E.g. “find image of a triumphant woman”.
  - Requires very complex logic.
  - Far from being available at present level of technology.

Image retrieval



CBIR framework example



Naive per-pixel comparison

- Pixels are the most privitive features, so…
- Compare images on a per-pixel basis.
- Feature vector: raw array of pixel intensities.

$$D(I, Q) = \sum_r \sum_c d_c(I(r, c), Q(r, c)).$$

Bad Idea!

Why?q

Image/audio fingerprints

A fingerprint is a content-based compact signature that summarises some specific audio/video content.

Requirements:

- Discriminating power.
  - Ability to accurately identify an item within a huge number of other items (e.g. large audio collection in Shazam, millions of songs).
  - Low probability of false positives.
  - Query potentially has low information content: a few seconds of audio, a crude sketch of an image.

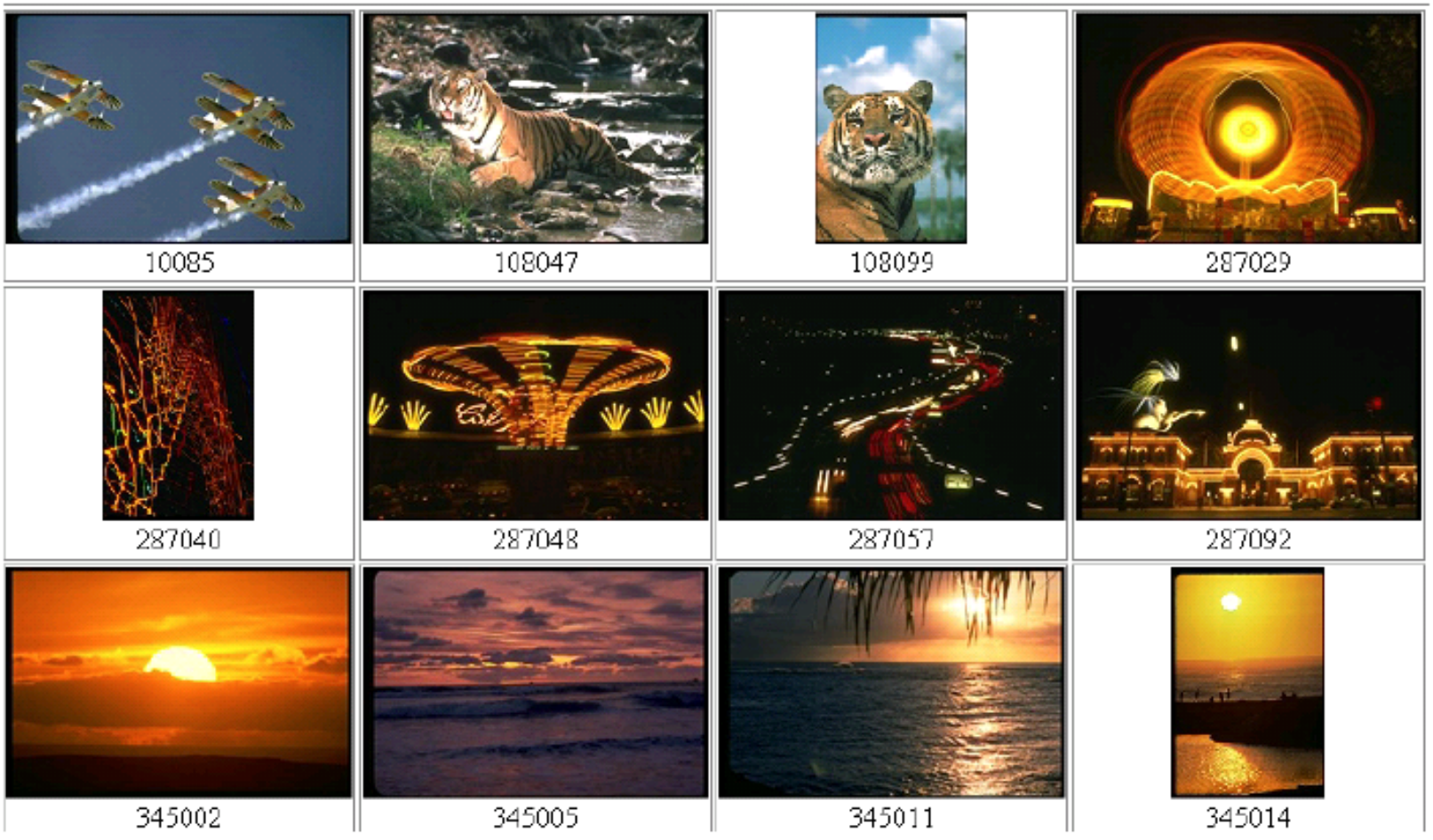
Image/audio fingerprints

- Invariance to distortions.
  - Shazam audio query may be distorted and superimposed with other audio sources.
  - Background noise.
  - Transformations: image rotation/scale/translation, warping. Lighting variations. Audio may be played faster or slower.
  - Compression artifacts
  - Cropping, framing.
- Compactness.
  - Making indexing feasible.
  - Allowing for fast search.
- Computational simplicity.
  - E.g. for use on mobile devices.

Feature extraction in images

- Object identification, e.g.
  - Detect faces (realatively robust these days).
  - Segmentation into blobs.
  - Text detection/OCR.
  - General case is difficult.
- Colour statistics, e.g. histogram (3-dimensional array that counts pixels with specific RGB or HSV values in an image.)
- Colour layout, e.g. “blue on top, green below”.
- Texture properties, usually based on edges in image.
- Motion information (in videos).

Search by colour histogram



Search by colour histogram of sunset
(scores shown under images).

Histogram comparison

- For each i-th training image generate colour histogram H_d.
- Normalise it so that is sums to one (to reduce the effect of the size of image).
- Store it as the feature in the database.
- For a query image, also compute histogram H_q.

Histogram comparison

- Compare against the database using histogram intersection:

$$\text{Intersection} = \sum_i \min(H_d^i, H_q^i).$$

For similar histograms (images) the intersection is closer to 1.

- Another standard measure of similarity for color histograms:

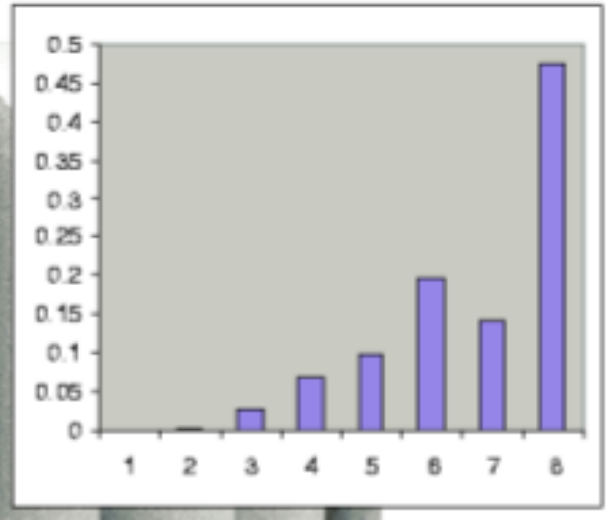
$$\text{Difference} = (H_d - H_q)^T A (H_d - H_q),$$

where A is a similarity matrix.

- Or simply L_1 norm:

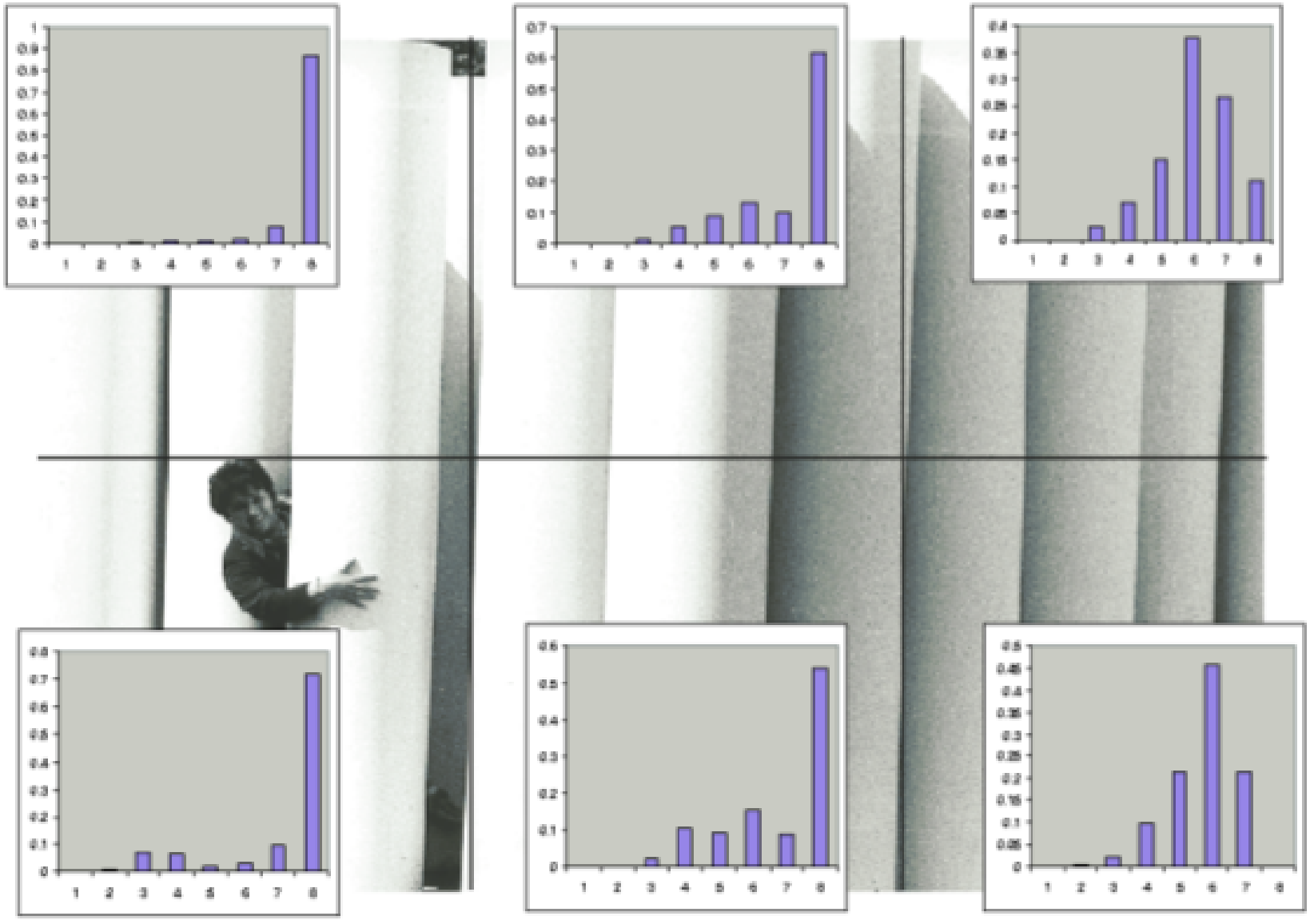
$$\text{Difference} = \sum |H_d^i - H_q^i|.$$

Search by colour histogram



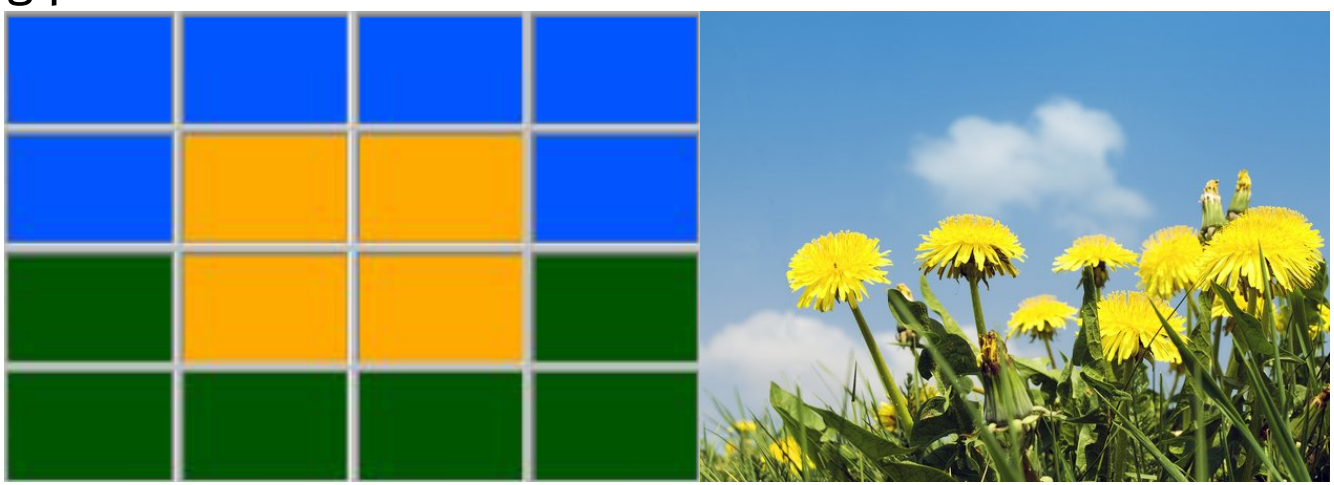
1:	0	–	31
2:	32	–	63
3:	64	–	95
4:	96	–	127
5:	128	–	159
6:	160	–	191
7:	192	–	223
8:	224	–	255

Search by colour histogram

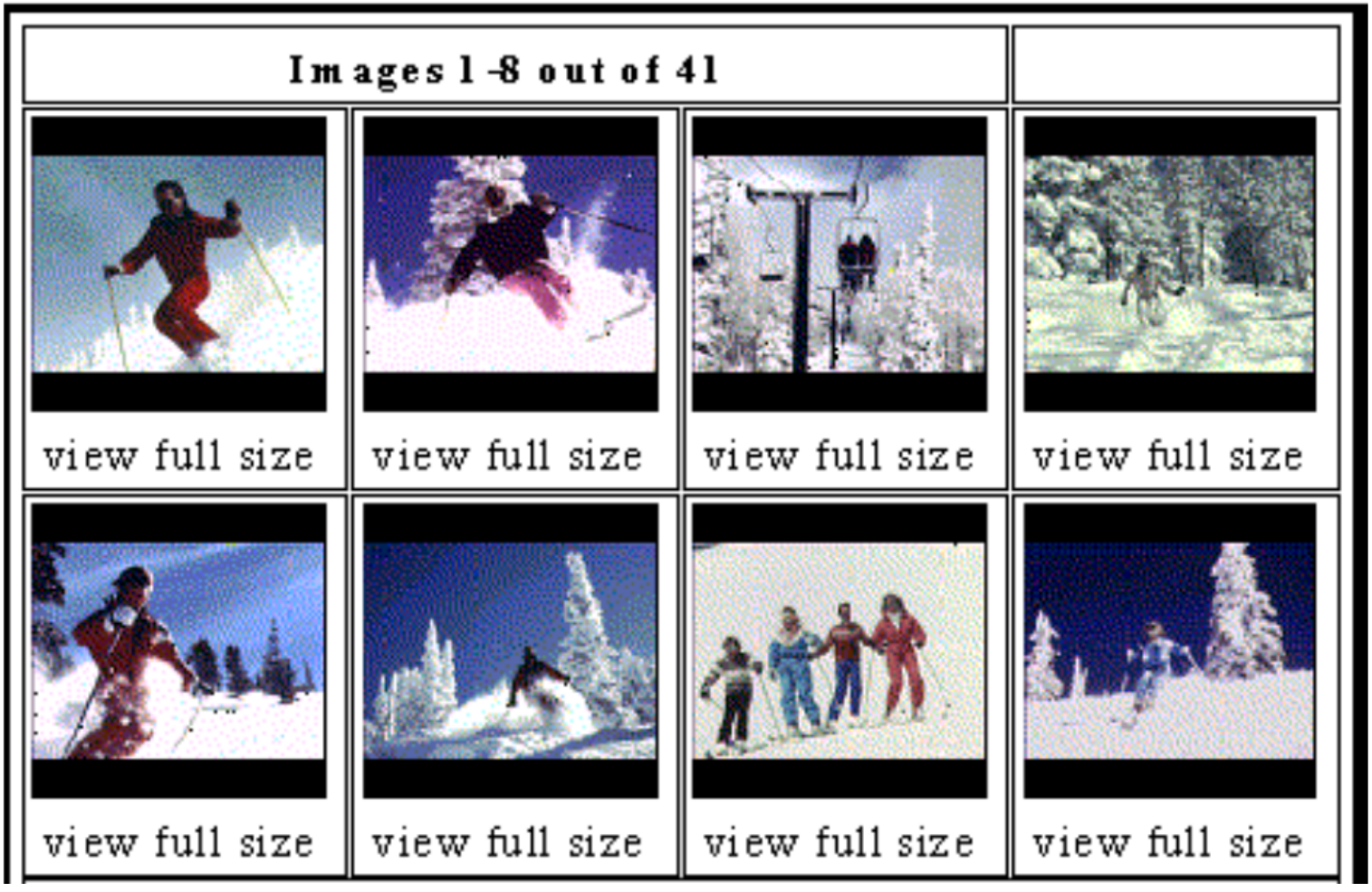


Search by colour layout

- An improvement over basic colour/histogram search.
- The user can set up a scheme of how colors should appear in the image, in terms of coarse blocks of colour, e.g. on a grid.
- The training images are partitioned into regions and histograms (or simply average colours) are computed for each region.
- Matching process is similar.



Search by colour layout

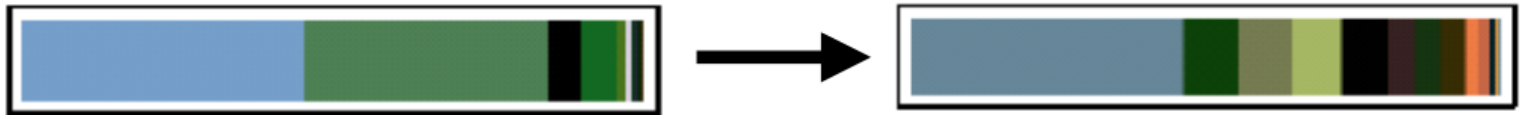


Retrieval by “color layout” in IBM’s QBIC system.

Colour signatures and EMD

For each image, compute color signature:

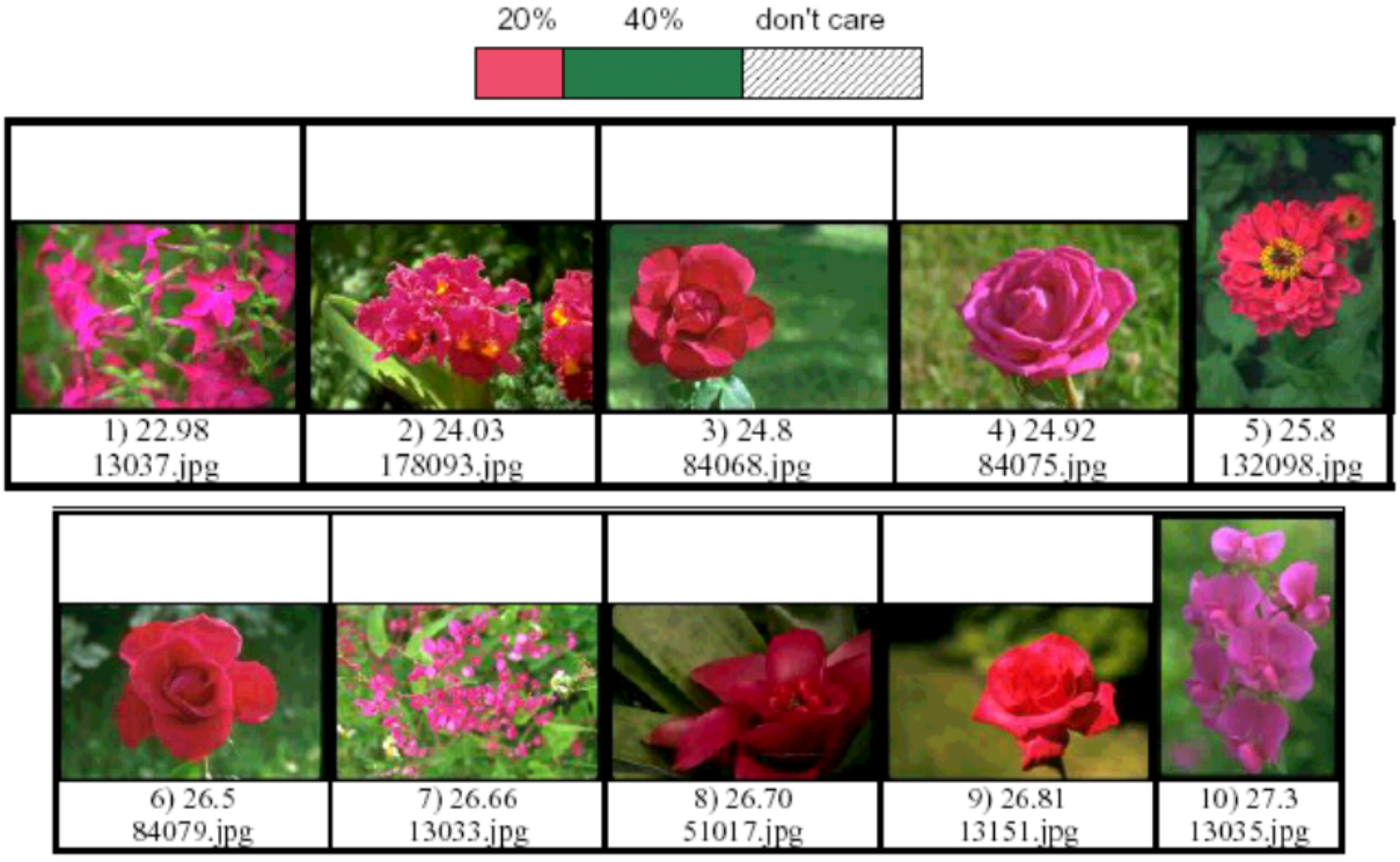
Define distance between two color signatures to be the minimum amount of “work” needed to transform one signature into another (earth mover’s distance):



Colour signatures and EMD

- Transform pixel colors into CIE-LAB color space.
- Each pixel of the image constitutes a point in this color space.
- Cluster the pixels in color space. (Clusters constrained to not exceed R units in L,a,b axes.)
- Find centroids of each cluster.
- Each cluster contributes a pair (μ, w) to the signature.
- μ is the average color.
- w is the fraction of pixels in that cluster.
- Typically there are 8 to 12 clusters.

Colour signatures and EMD



[Rubner, Guibas, & Tomasi 1998]

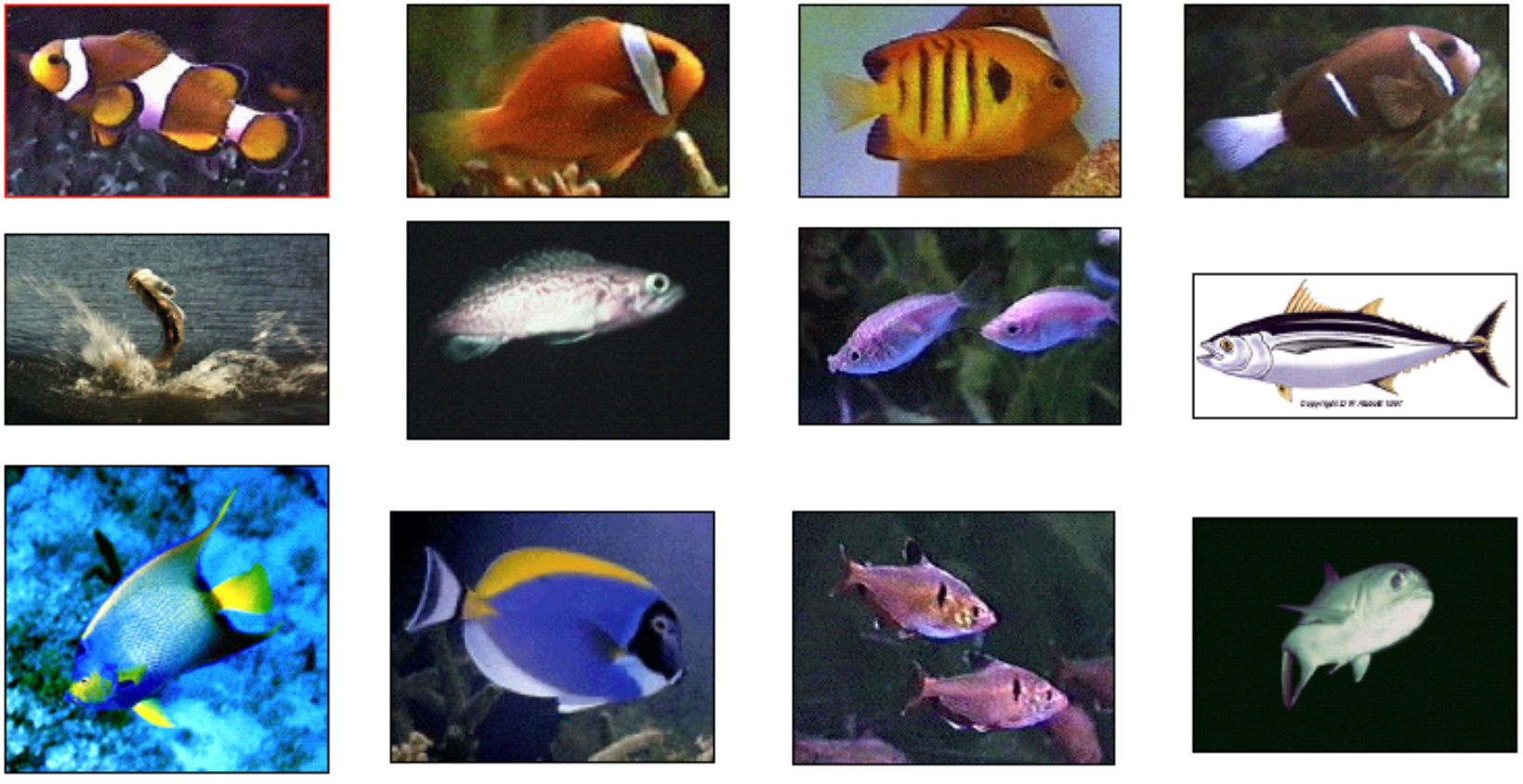
Visualisation using MDS with EMD as Distance

[Rubner, Guibas, & Tomasi 1998]

Search by sketch

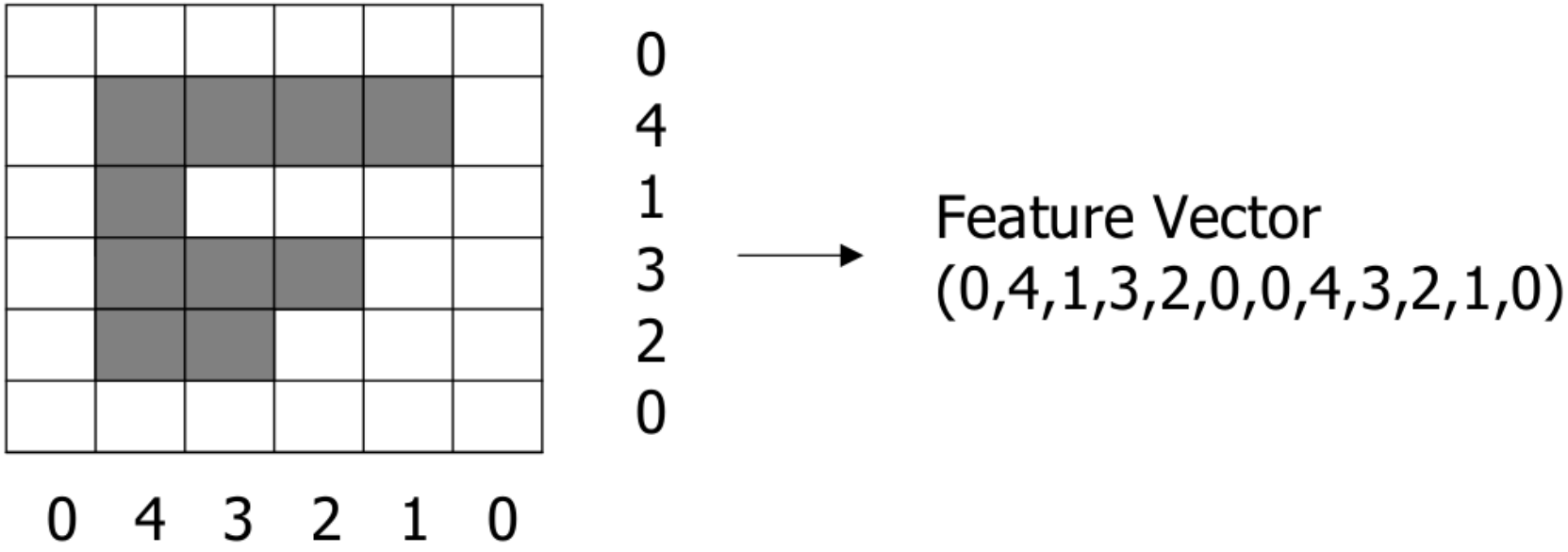
Example taken from Jacobs, Finkelstein, & Salesin
Fast Multi-Resolution Image Querying, SIGGRAPH 1995

Search by shape



(Query shape in top left corner.)

Projection matching



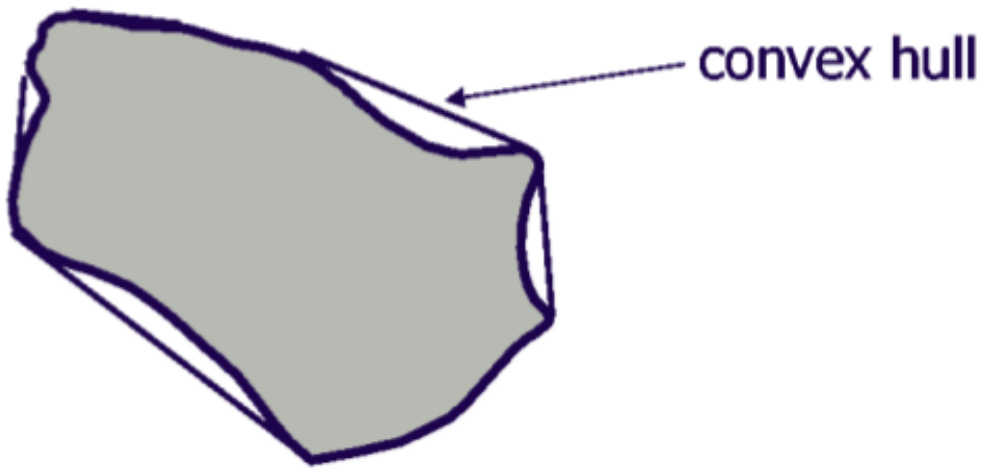
[Smith & Chang, 1996]

In projection matching, the horizontal and vertical projections of a shape silhouette form a histogram.

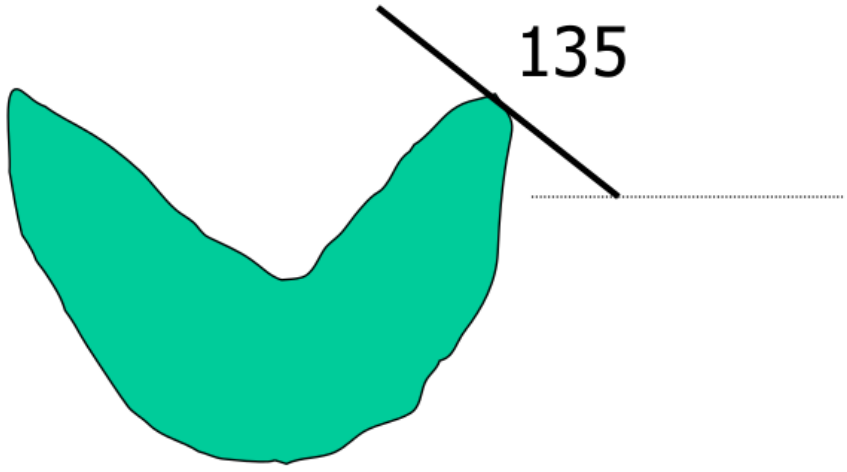
- Weaknesses?
- Strengths?

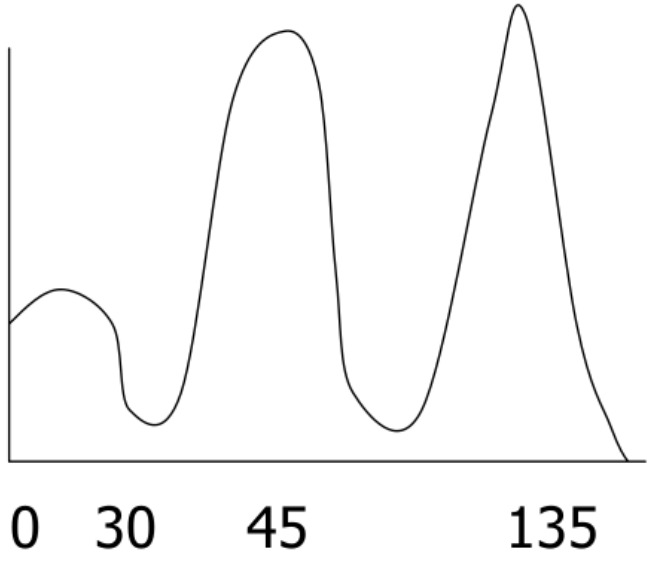
Area and perimeter

- Circularity (compactness): $C = 4\pi \frac{A}{P^2}$.
  - C is 1 for circle, smaller for other shapes.
- Convexity: ratio of perimeter of convex hull and original curve.

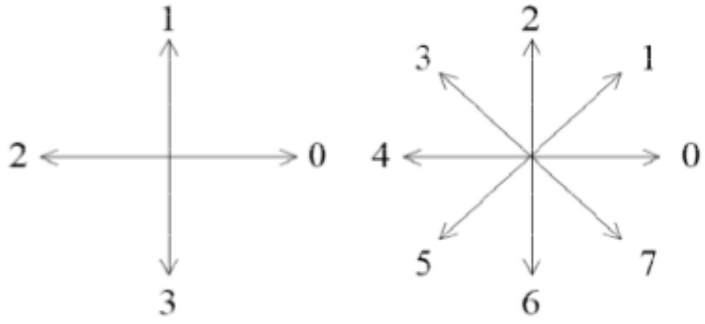


Tangent angle histograms

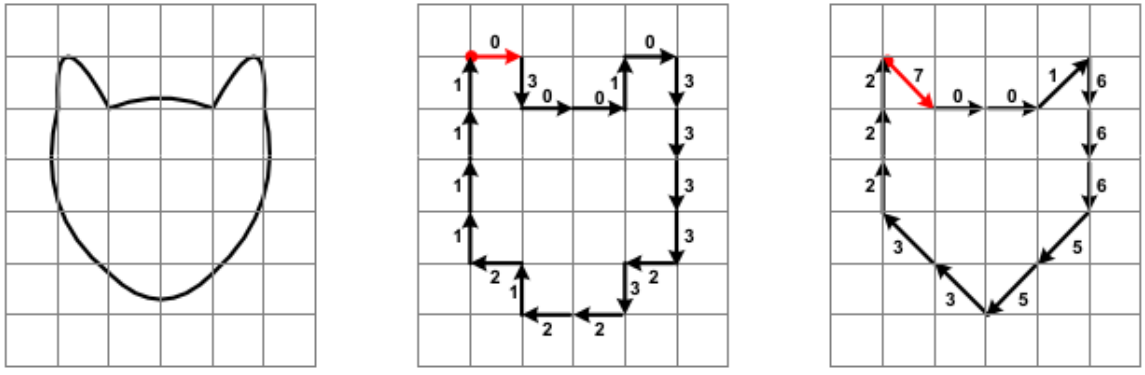




Chain codes

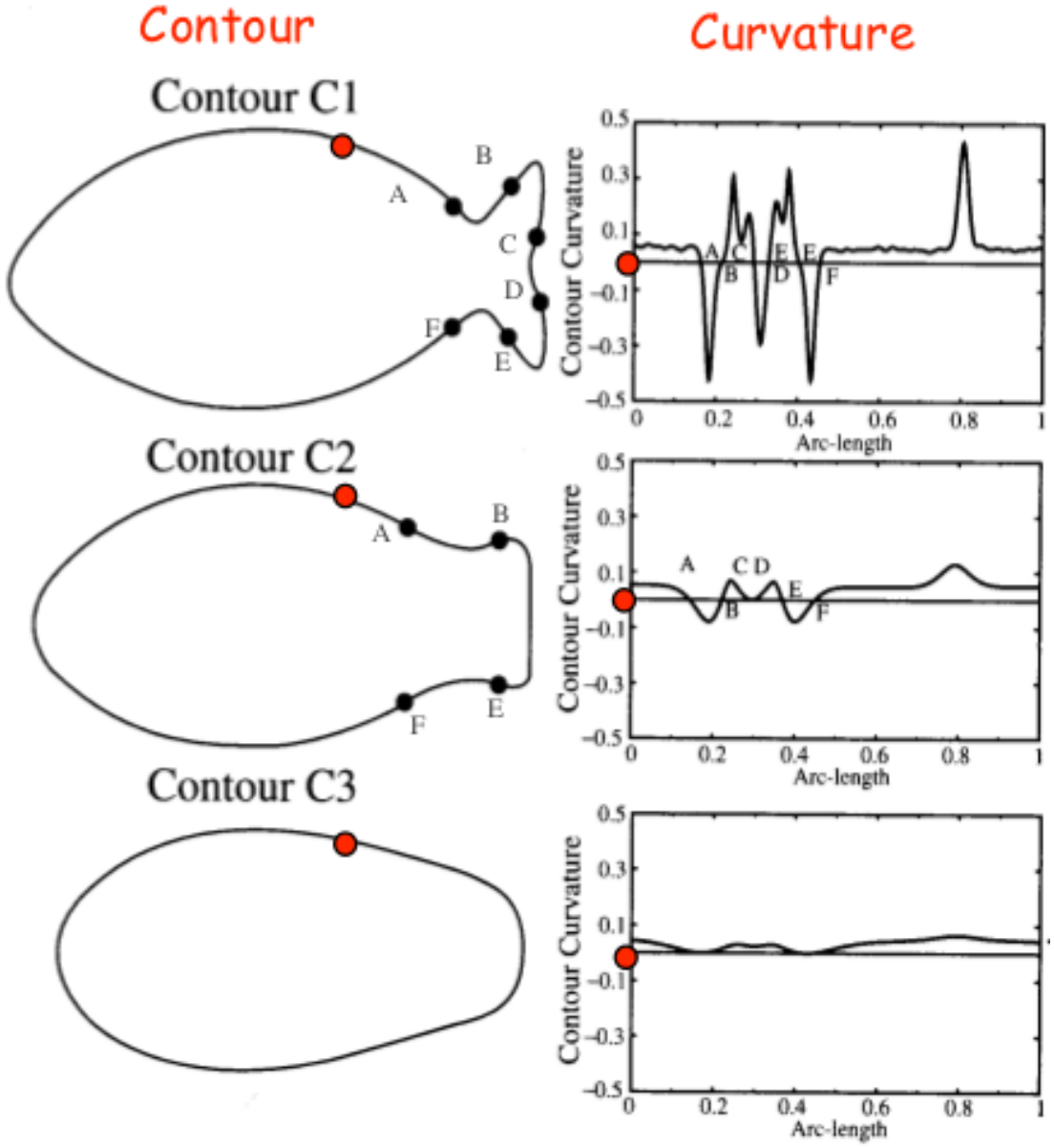


Example:



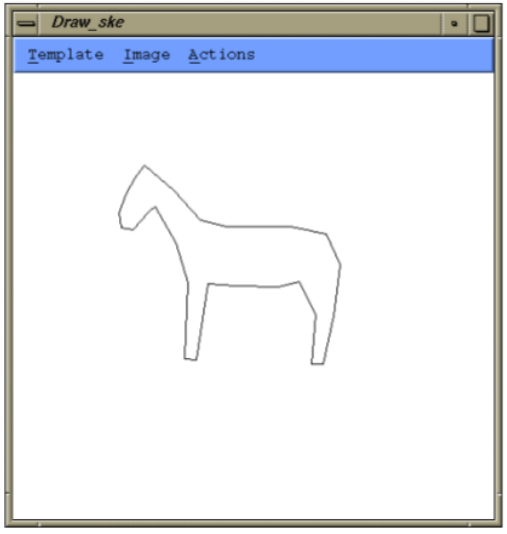
- Sorting chain codes makes them invariant to starting point.
- Use histograms of chain codes.

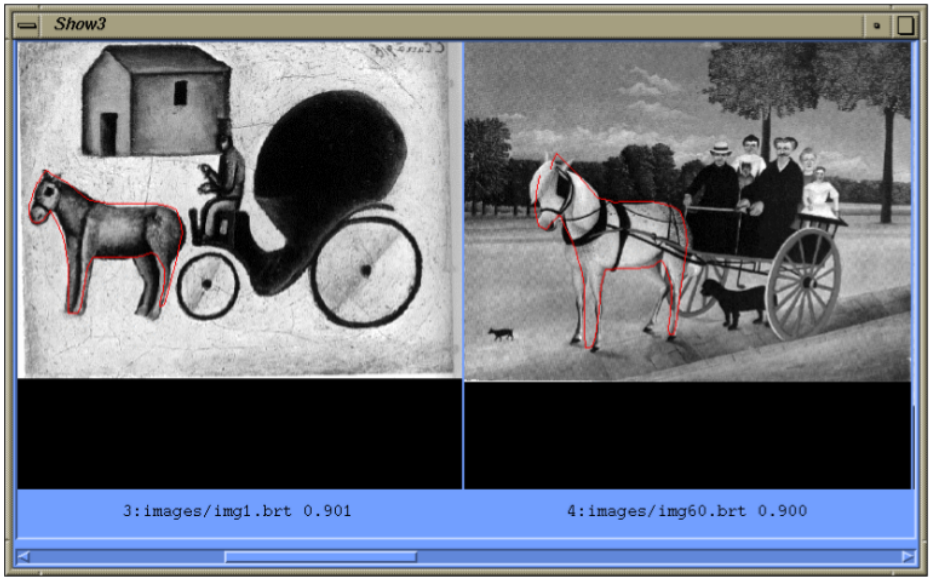
Curvature

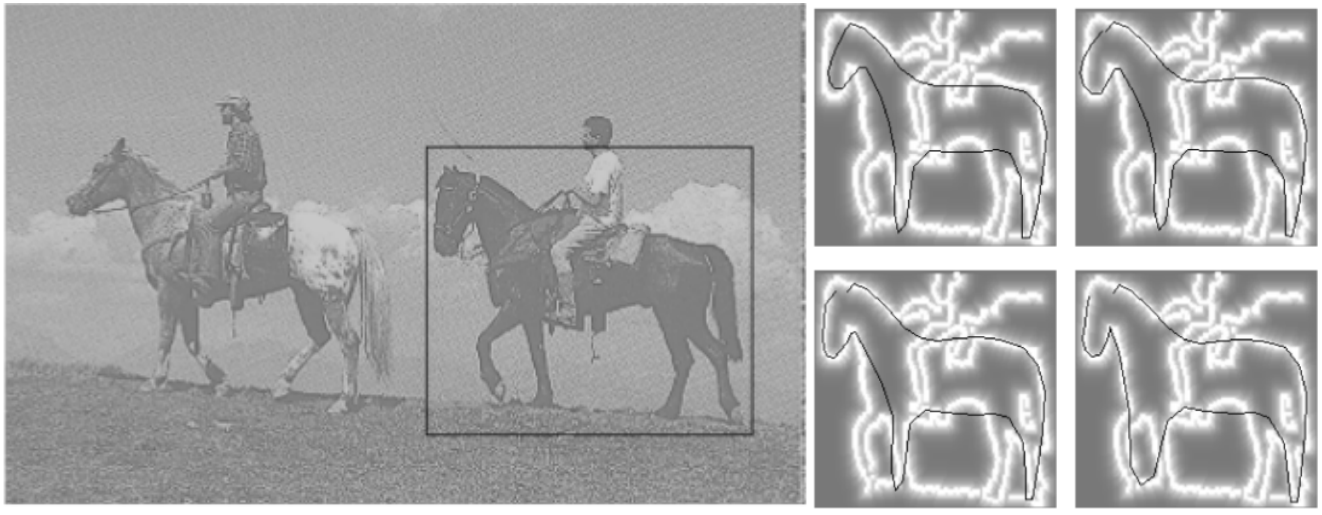

Contour
Curvature
Contour C1
A
B
C
D
E
F
Contour Curvature
Arc-length
Contour C2
A
B
E
F
Contour Curvature
Arc-length
Contour C3
Contour Curvature
Arc-length

Elastic shape matching

[Del Bimbo & Pala, 1997]





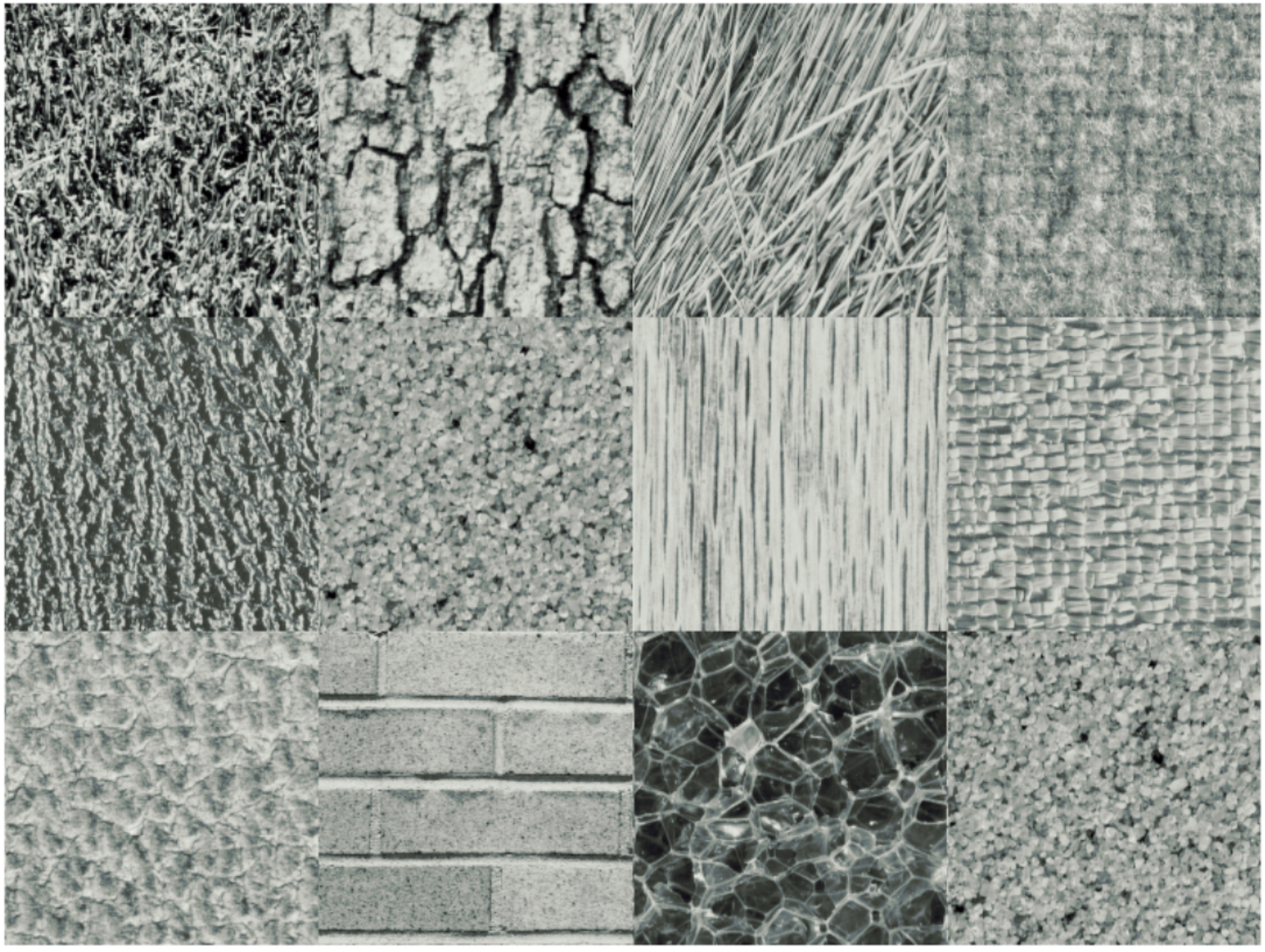


Shape matching problems

Many existing shape matching approaches assume

- Segmentation is given.
- Human selects object of interest.
- Lack of clutter and shadows.
- Objects are rigid.
- Planar (2-D) shape models.
- Models are known in advance.

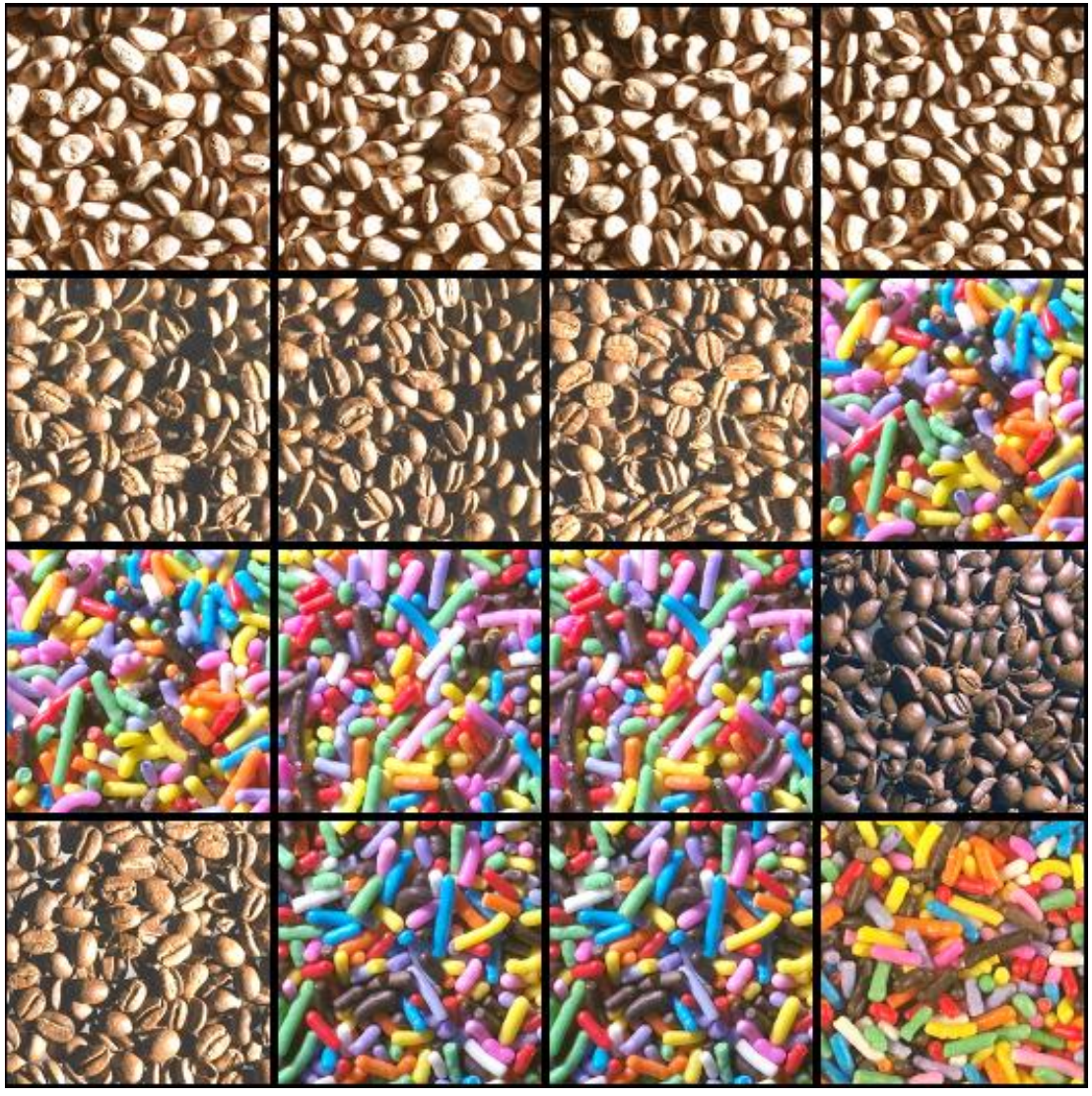
Texture



Texture

- Texture is a perceptual phenomenon due to local variations in image intensity.
- Local region property.
- Less local than pixel, more local than objects/entire image.
- Usually repeated pattern with salient statistical properties.

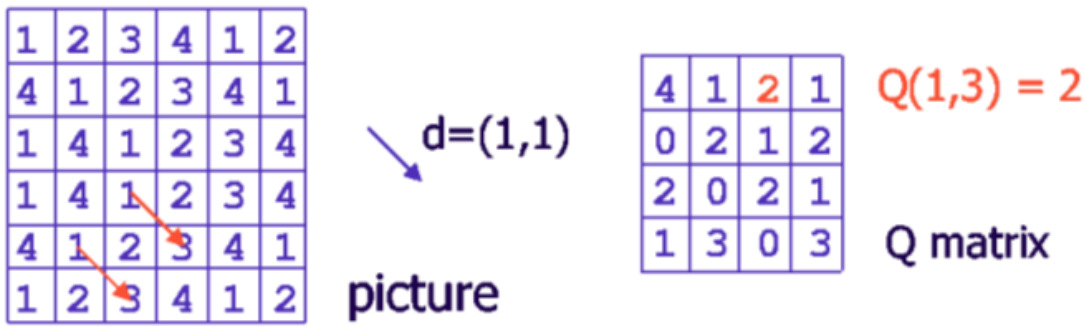
Search by texture



(Query shape in top left corner.)

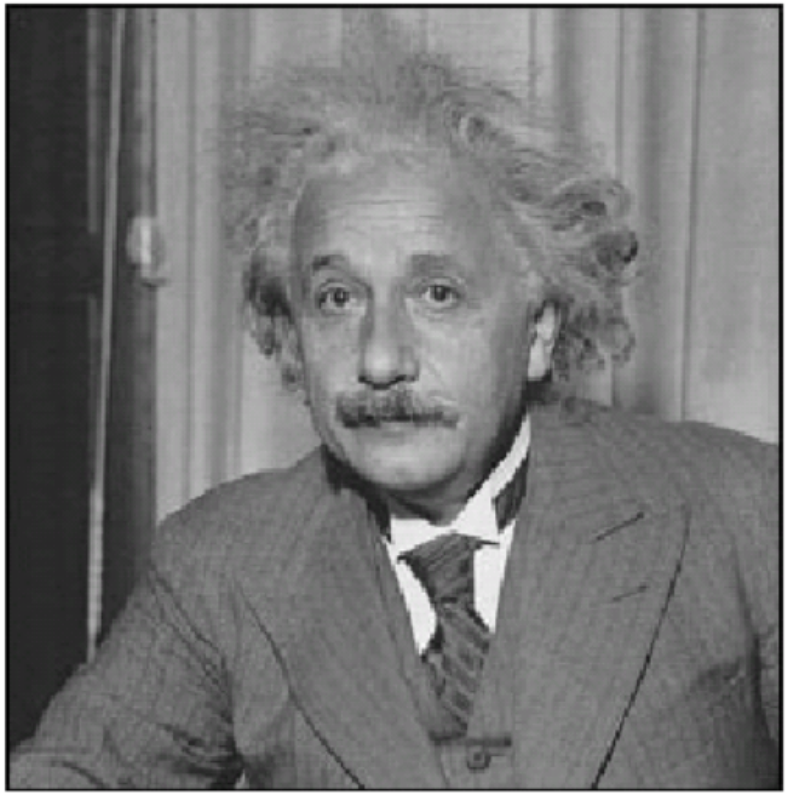
Co-occurence

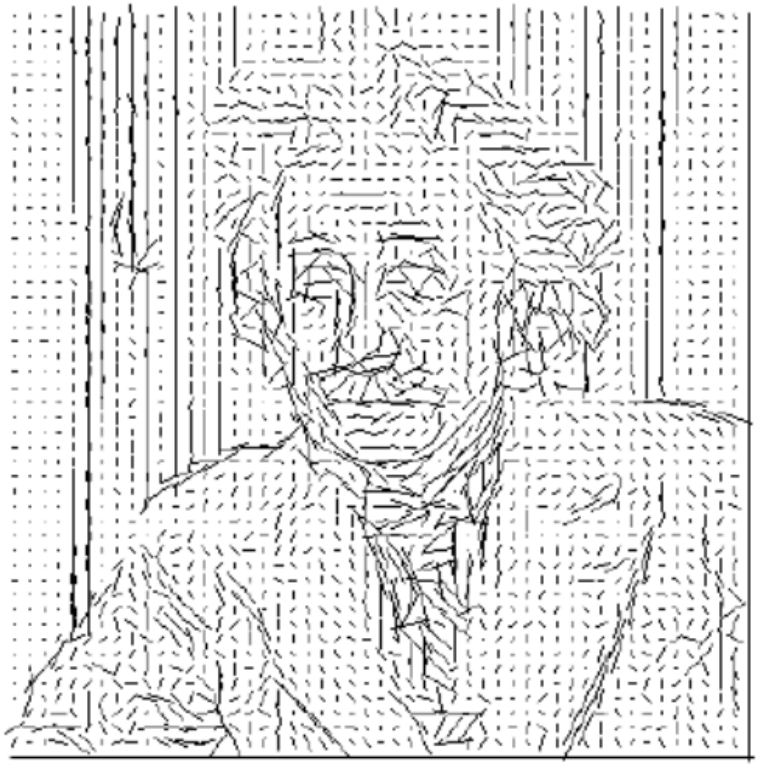
- We can capture some spatial properties of texture with co-occurence histogram.
- For a displacement vector $\boldsymbol{d} = (d_x, d_y)$:
- Count in $N \times N$ bins of $Q(i, j)$ how many times gray levels i and j are separated by displacement $\boldsymbol{d}$ in the image.
- Q captures some spatial information about distribution of gray levels.
- Statistical properties: entropy $-\sum Q(i, j) \log Q(i, j)$, energy $\sum Q^2(i, j)$, contrast $\sum (i - j)^2 Q(i, j)$.



Orientation histograms

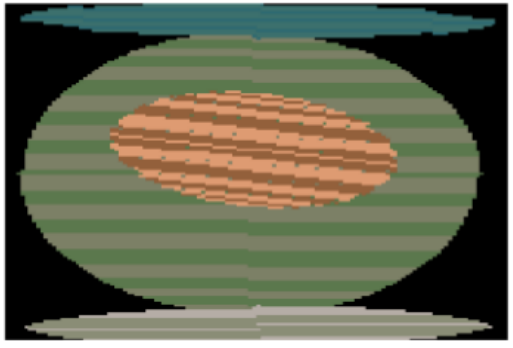
Determine local orientation and magnitude at each pixel:





If magnitude greater than threshold, increment corresponding histogram bin. [Freeman & Adelson, 1991]

Blobworld


- Images are segmented on colour plus texture.
- User selects a region of the query image.
- System returns images with similar regions.

Blobworld

Search by text

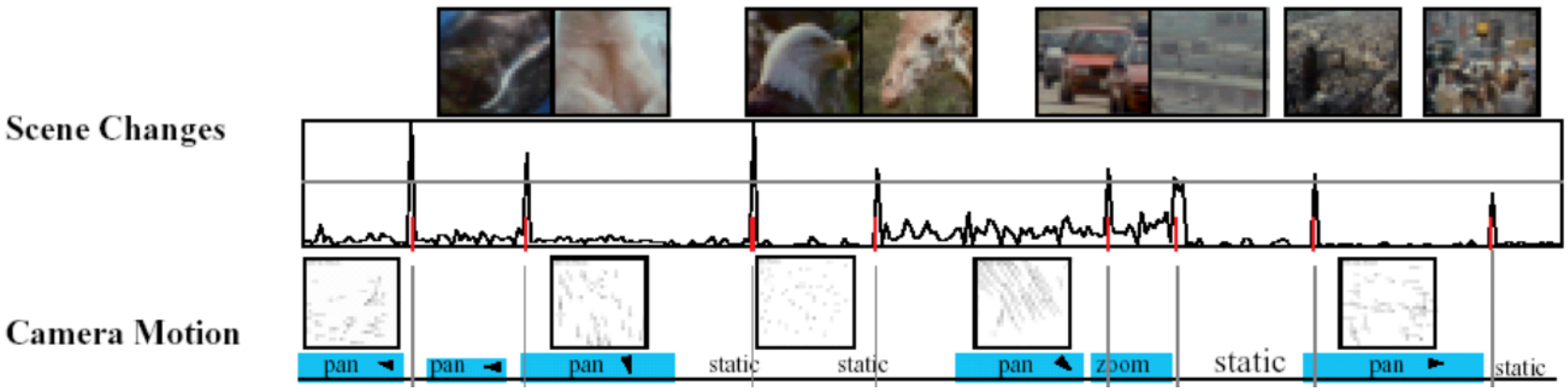
Parse text, essentially reducing the problem to traditional search.

Representative frames in videos

- Shots are a sequence of contiguous video frames grouped together:
  - Same scene.
  - Single camera operation.
  - Significant event.
- Automatic shot boundary detection:
  - Change in global color/intensity histogram.
  - Camera operations like zoom and pan.
  - Change in object motion.
- Representative frames:
  - Video broken into shots, and representative frames are selected.
  - Reduce video retrieval problem to image retrieval.
  - E.g. first, last, middle.

Representative frames in videos

Representative frames in videos



Content-based audio retrieval

Example scenarios:

- Song stuck in the head:
  - Search by humming.
  - Search by notes, contour, rhythm. E.g. Musipedia.
- What song is playing now? Search by audio
  e.g. Shazam.

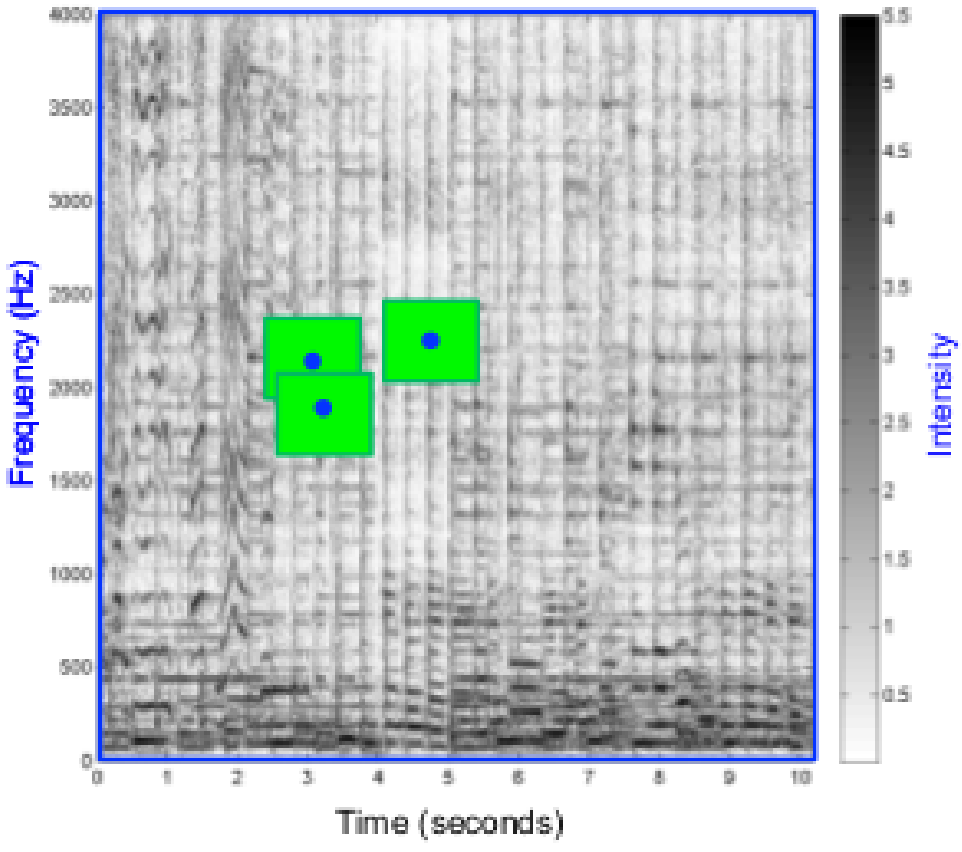
Audio search: how Shazam works

- Off-line: a large database of audio recordings (in feature space).
  - If metadata available then it is possible to name title, artist etc.
- Query: short audio fragment (5–15 sec). Mobile phone = low quality.
- Goal: identify recording where audio fragment came from.

Shazam fingerprints

- Experimentation revealed that spectrogram peaks is a good feature:
    - Robust to noise, room reverb, equalisation, overlapping sounds.
- A time-frequency point is a candidate peak if it has a higher energy content than all its neighbours in a region centered around the point.
- Density: make sure the entire audio covered approximately evenly.
- Choose peaks with higher amplitude. Reason: they are likelier to survive superposition of another sound.
    - Amplitude itself is not part of the fingerprint.

Shazam fingerprints (from Müller-Serrà paper)

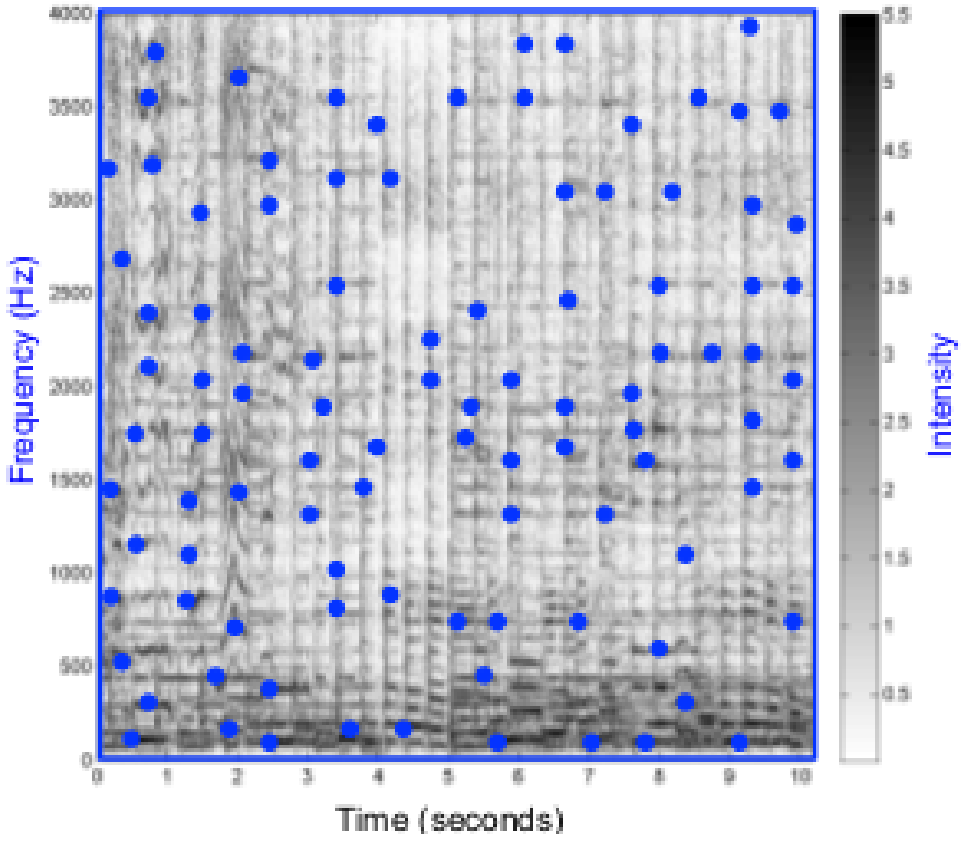


Steps:

1. Spectrogram
2. Peaks
   (local maxima)

- *Efficiently computable*
- *Standard transform*
- *Robust*

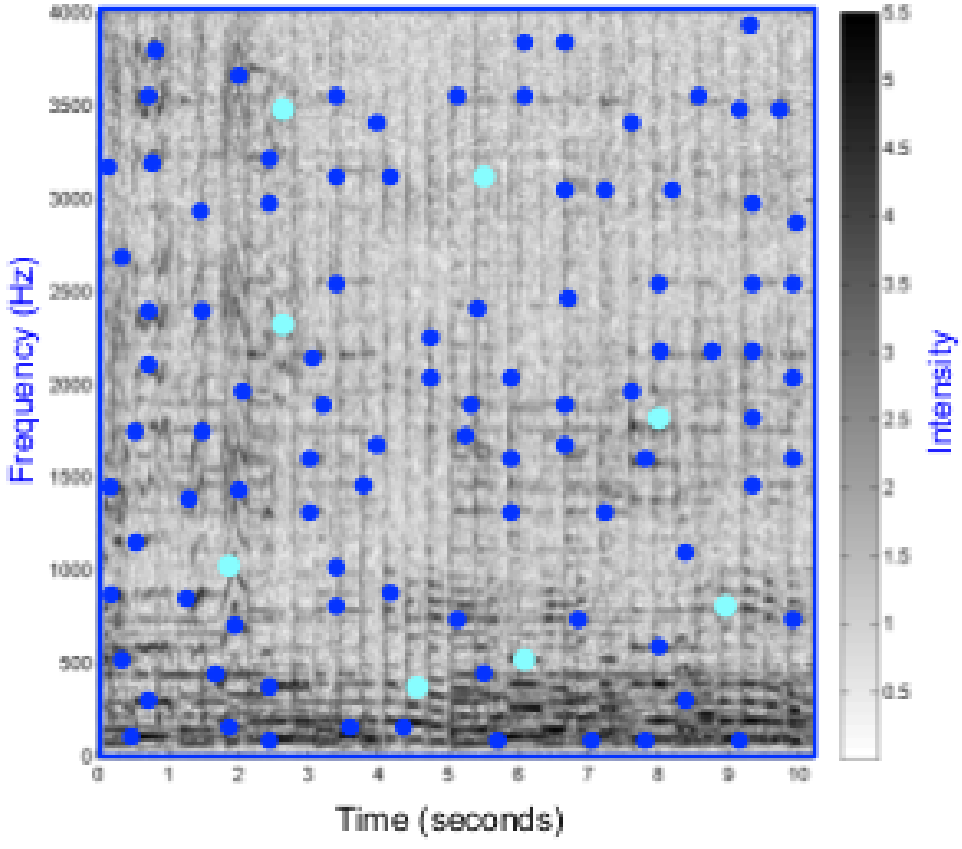
Shazam fingerprints



Steps:

1. Spectrogram
2. Peaks

Shazam fingerprints



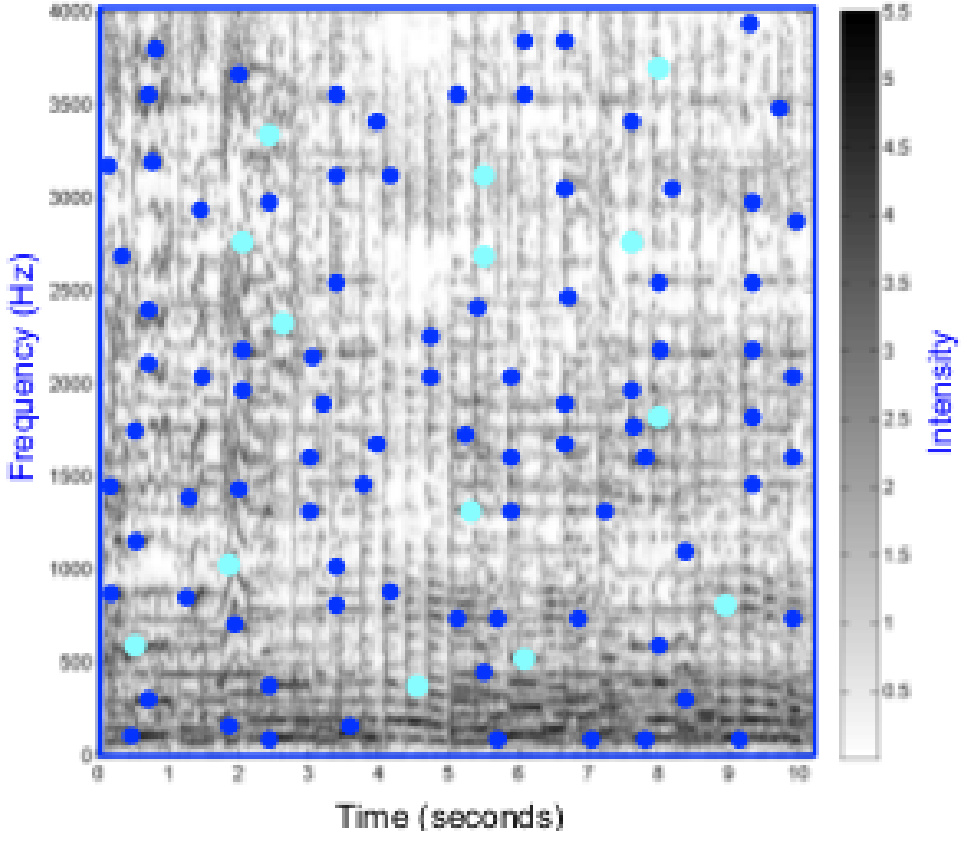
Steps:

1. Spectrogram
2. Peaks / differing peaks

Robustness:

- Noise, reverb, room acoustics, equalization

Shazam fingerprints



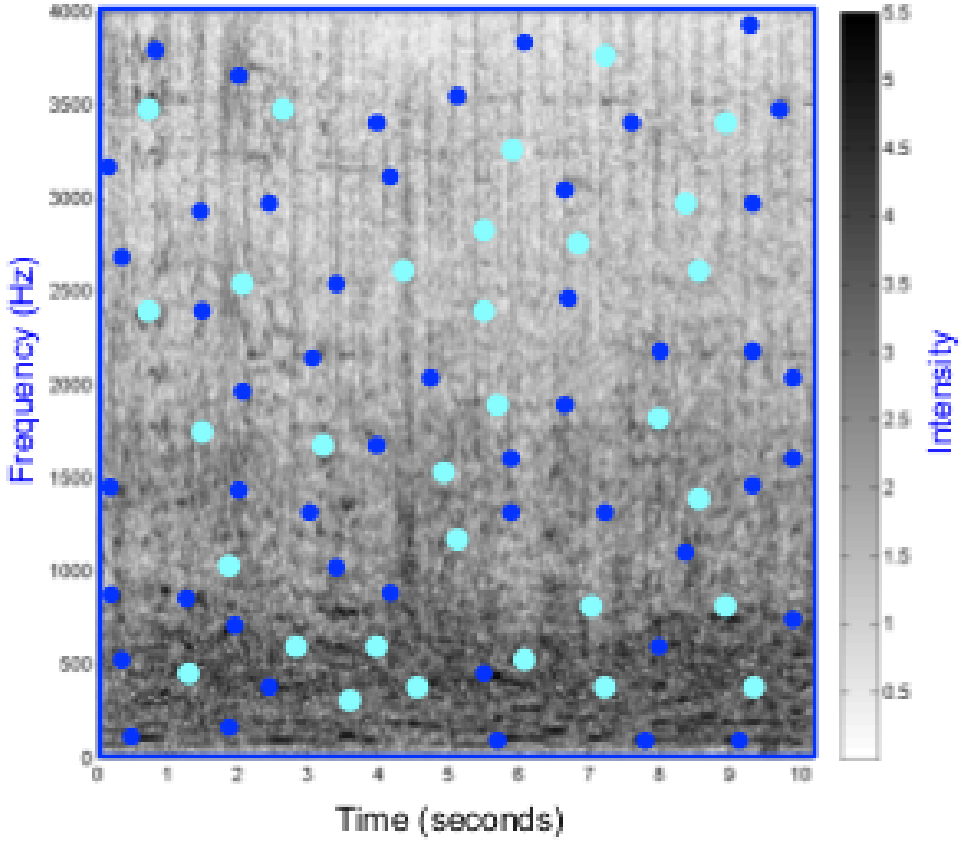
Steps:

1. Spectrogram
2. Peaks / differing peaks

Robustness:

- Noise, reverb, room acoustics, equalization
- Audio codec

Shazam fingerprints



Steps:

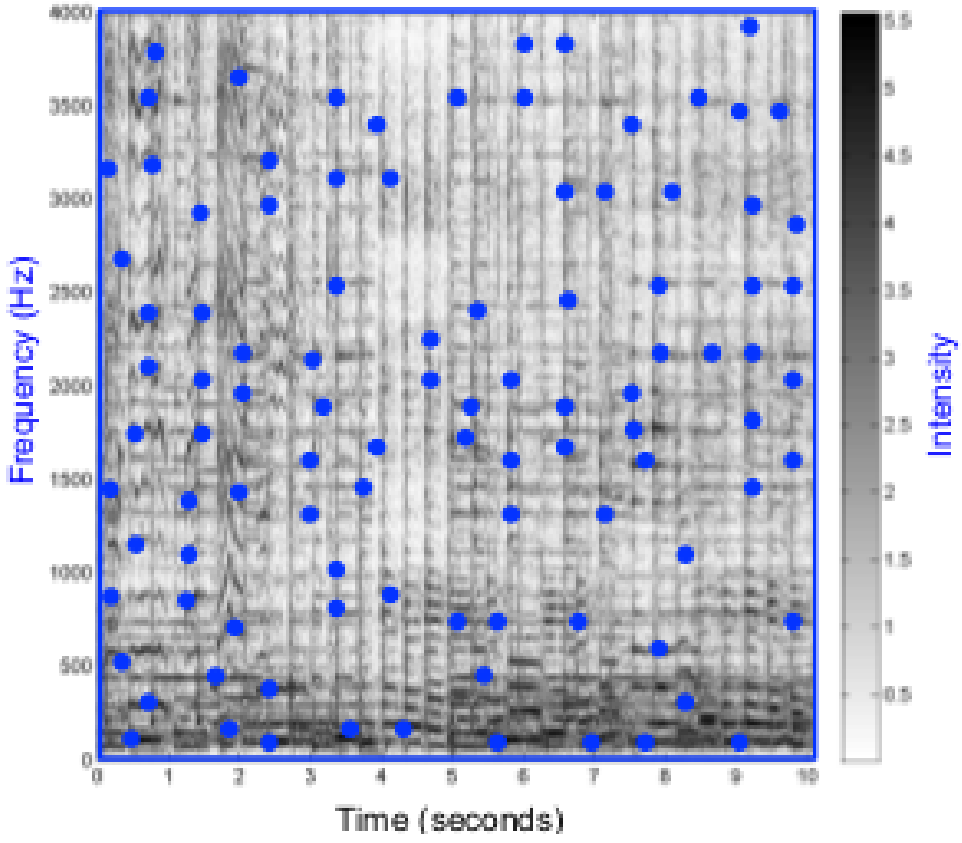
1. Spectrogram
2. Peaks / differing peaks

Robustness:

- Noise, reverb, room acoustics, equalization
- Audio codec
- Superposition of other audio sources

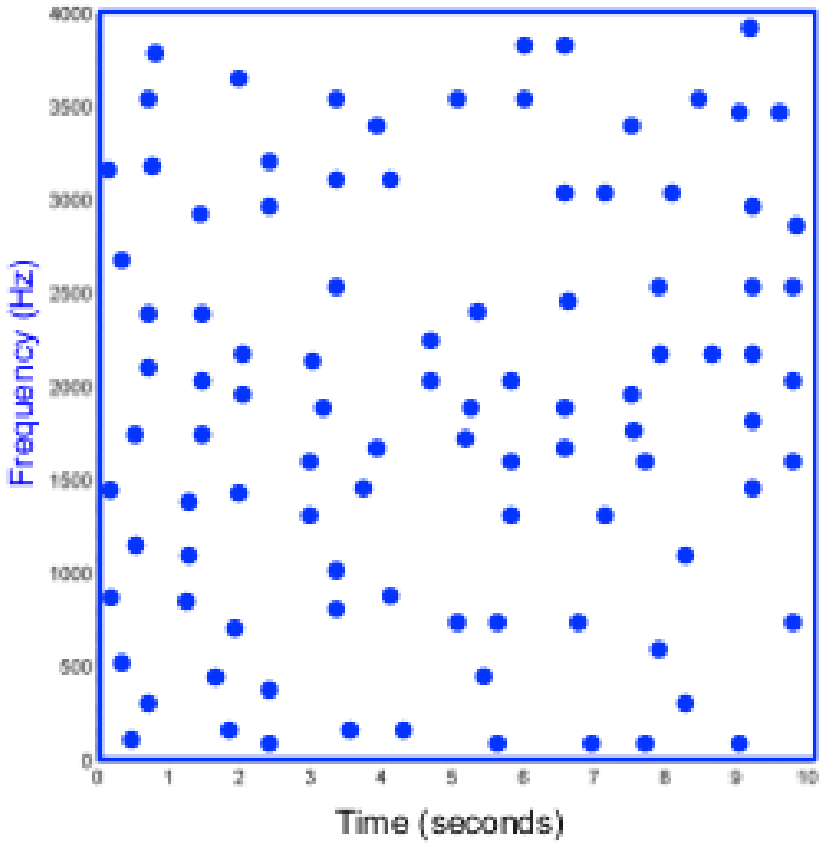
Shazam fingerprints

Database document

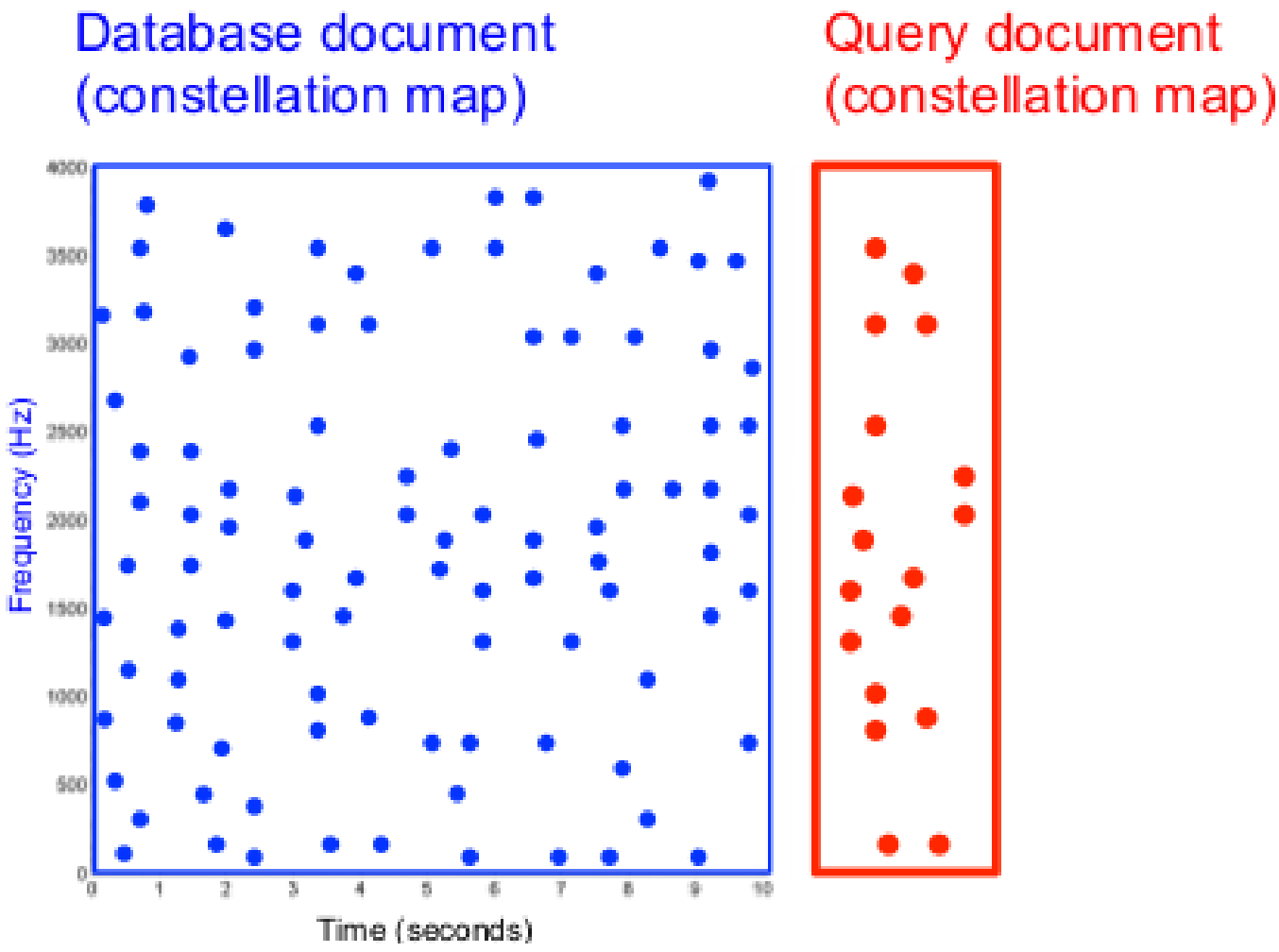


Shazam fingerprints

Database document
(constellation map)



Shazam fingerprints

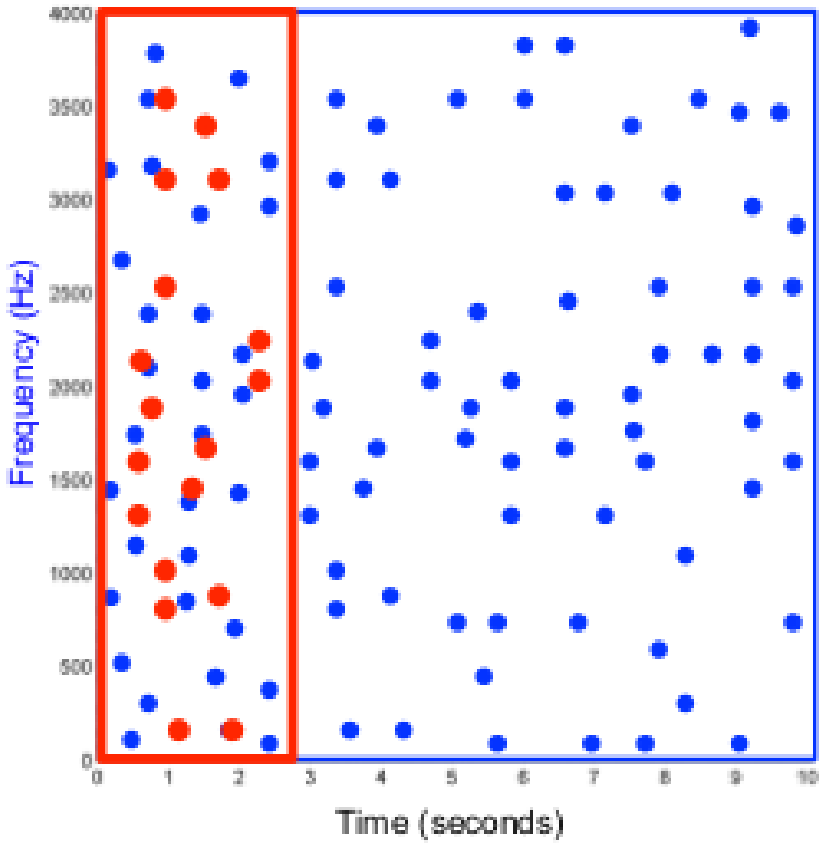


Shazam fingerprints

Database document
(constellation map)

Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks

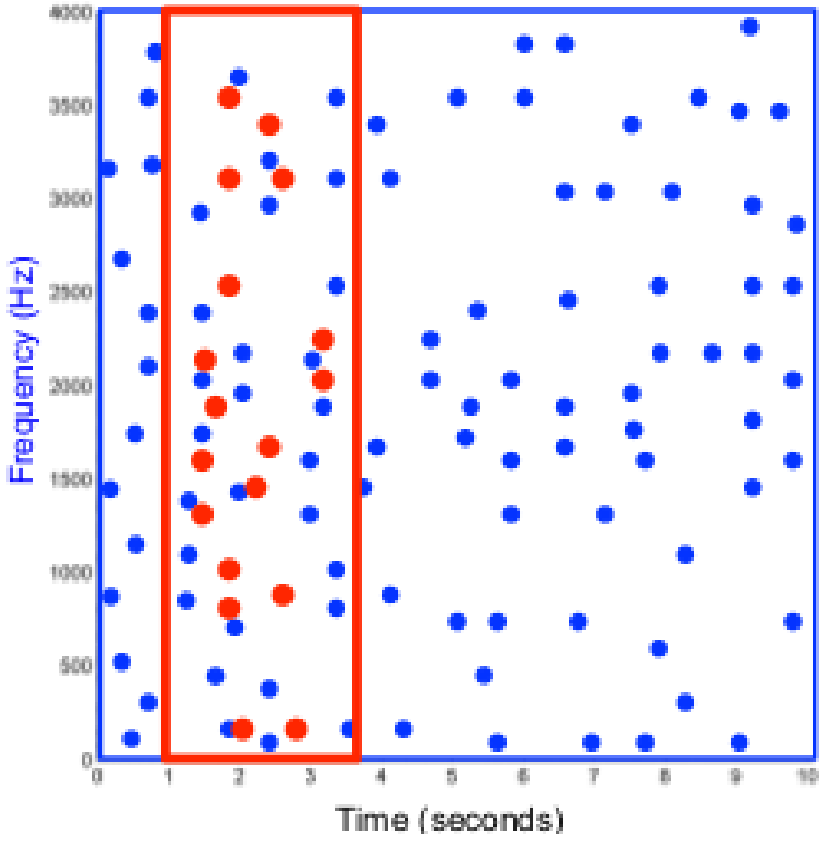


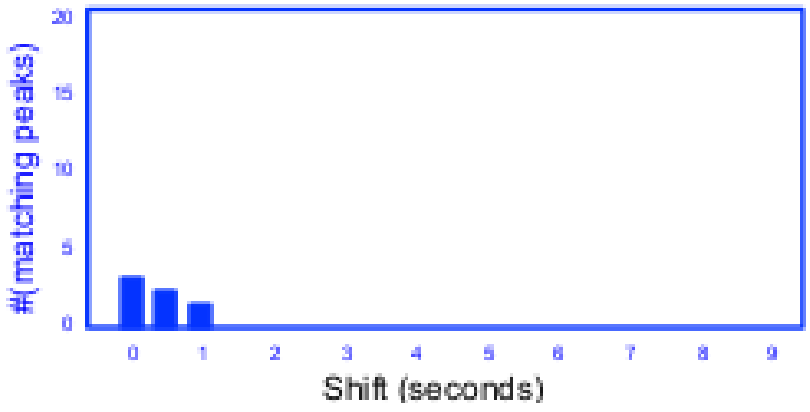
Shazam fingerprints

Database document
(constellation map)

Query document
(constellation map)

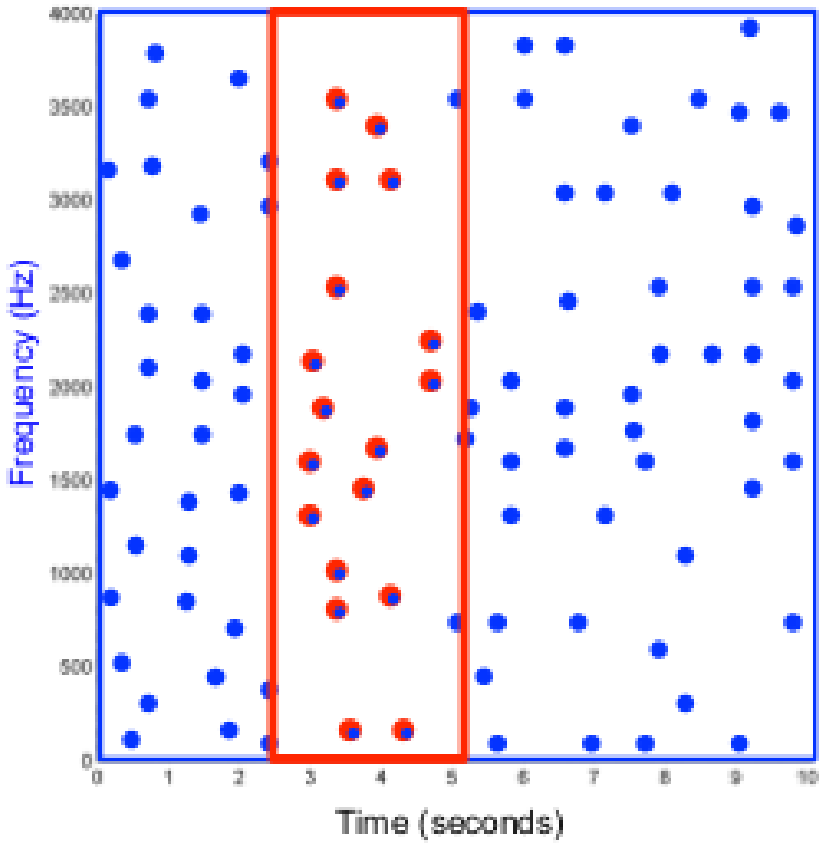
1. Shift query across database document
2. Count matching peaks





Shazam fingerprints

Database document
(constellation map)

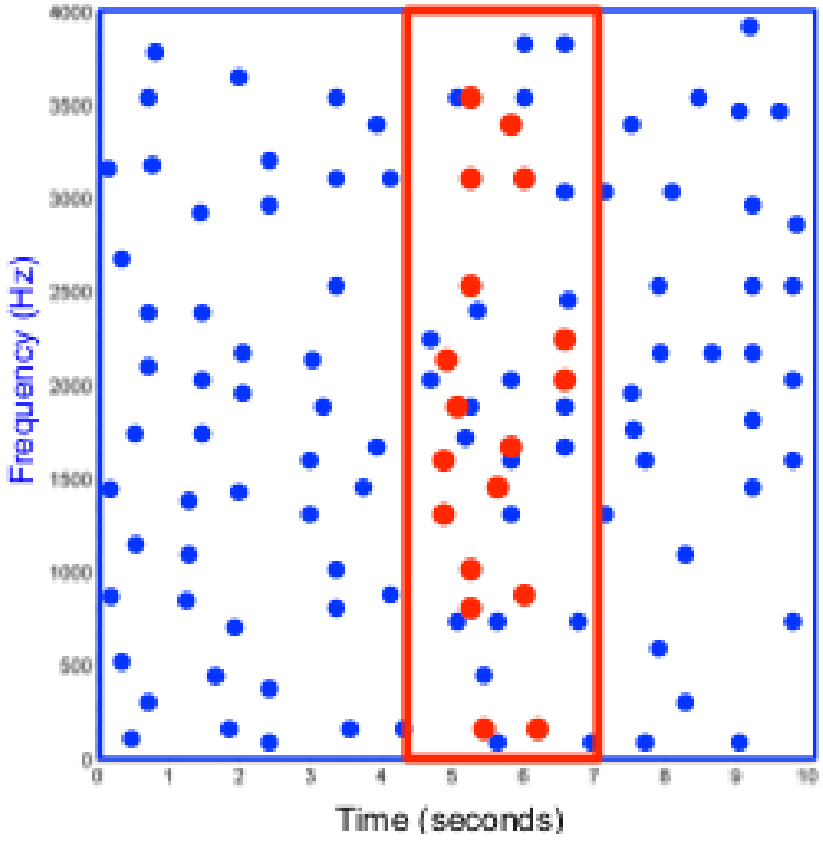


Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks

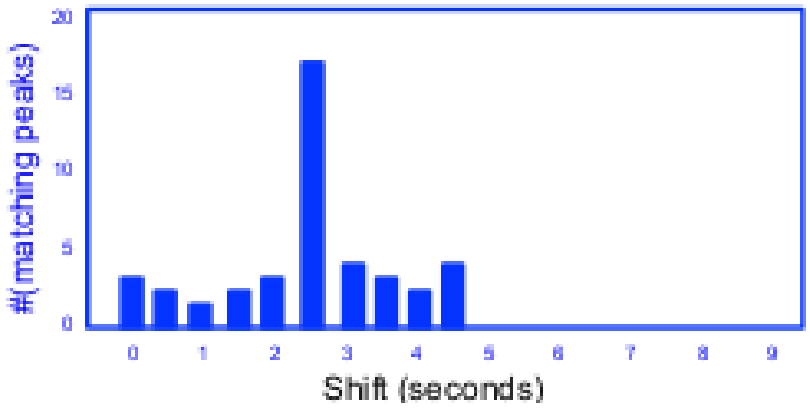
Shazam fingerprints

Database document
(constellation map)



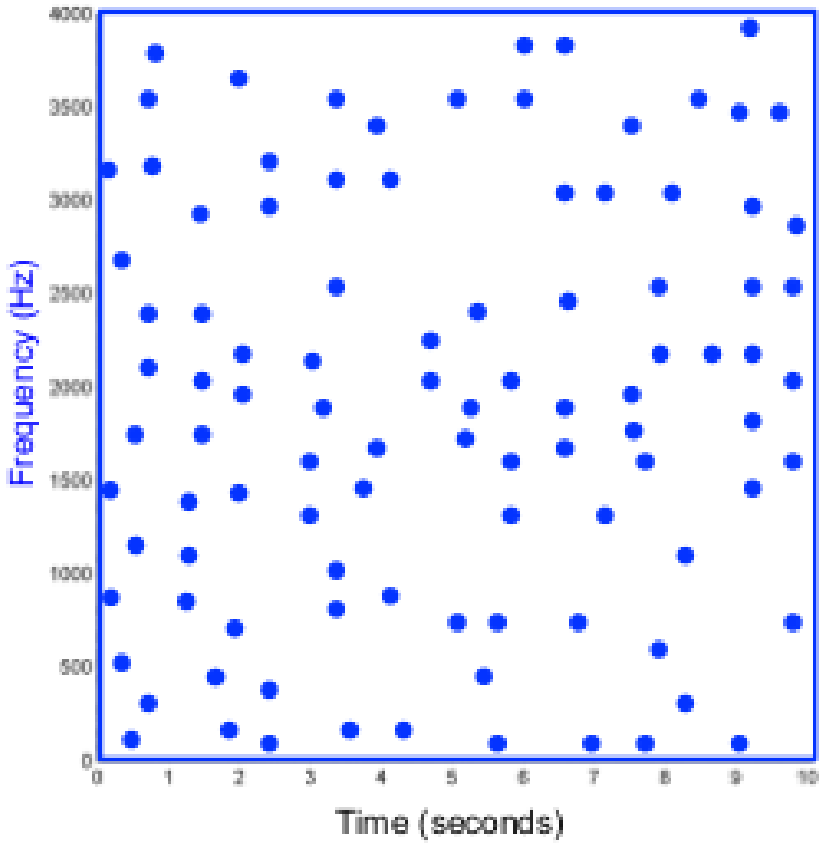
Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks



Shazam fingerprints

Database document
(constellation map)

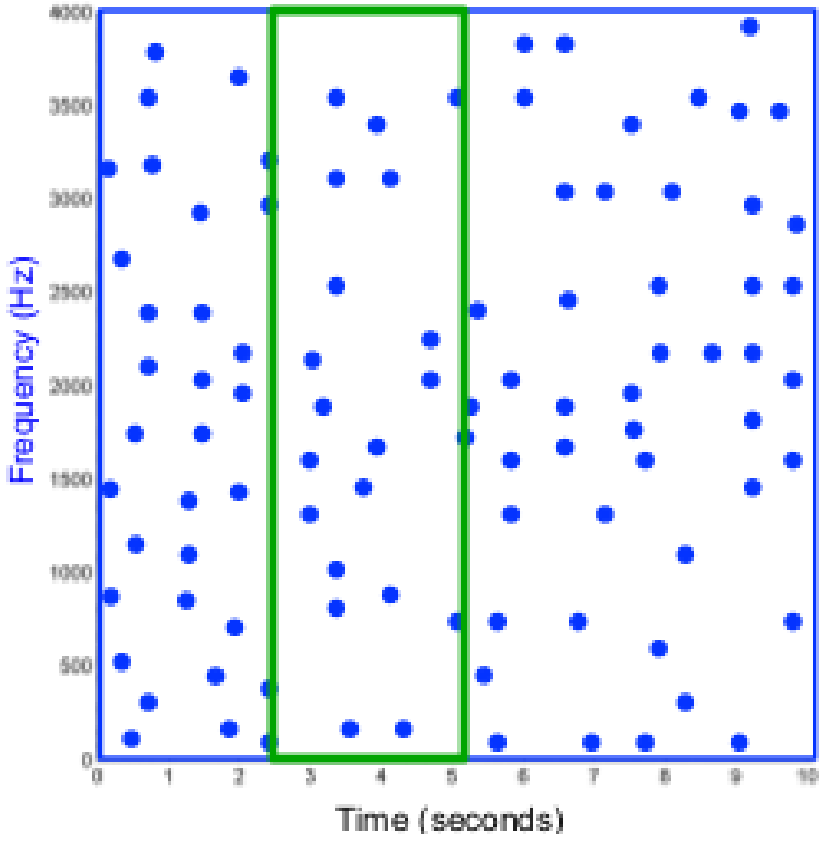


Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks

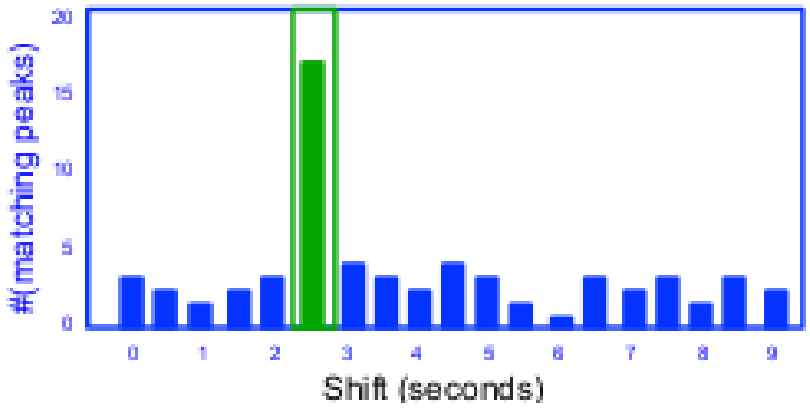
Shazam fingerprints

Database document
(constellation map)



Query document
(constellation map)

1. Shift query across database document
2. Count matching peaks
3. High count indicates a hit (document ID & position)



Further reading

- Original Shazam paper by Wang et al.
- Müller-Serrà paper on audio CBR of music.