Multimedia Information Retrieval

Prof Stefan Rüger

Multimedia and Information Systems
Knowledge Media Institute
The Open University

http://kmi.open.ac.uk/mmis

MMIS
Multimedia and Information Systems

KNOWLEDGE MEDIA
KMi
INSTITUTE

kmi.open.ac.uk

KMi

Home | News | Projects | Technologies | Publications | People | Seminars | Commercial

Working at KMi

Job Vacancies [2]

Studentships

Visitors / Internships

About KMi

Contact Us

Find Us

Events

Concept Videos

KMi Tools

Compendium

FM Technology

Magpie

About

The Knowledge Media Institute (KMi) was set up in 1995 in recognition of the need for the Open University to be at the forefront of research and development in a convergence of areas that impacted on the OU's very nature: Cognitive and Learning Sciences, Artificial Intelligence and Semantic Technologies, and Multimedia. We chose to call this convergence Knowledge Media.

Knowledge Media is about the processes of generating, understanding and sharing knowledge using several different media, as well as understanding how the use of different media shape these processes.

KMi operates with reference to a number of basic tenets, which define the context in which we formulate and pursue our research objectives:

Strategic Threads

Our research is aligned with a number of broad strategic threads, currently Future Internet, Knowledge Management, Multimedia & Information Systems, Narrative Hypermedia, New Media Systems, Semantic Web & Knowledge Services and Social Software.

Future Internet

Knowledge Management

Multimedia & Information Systems

Narrative Hypermedia

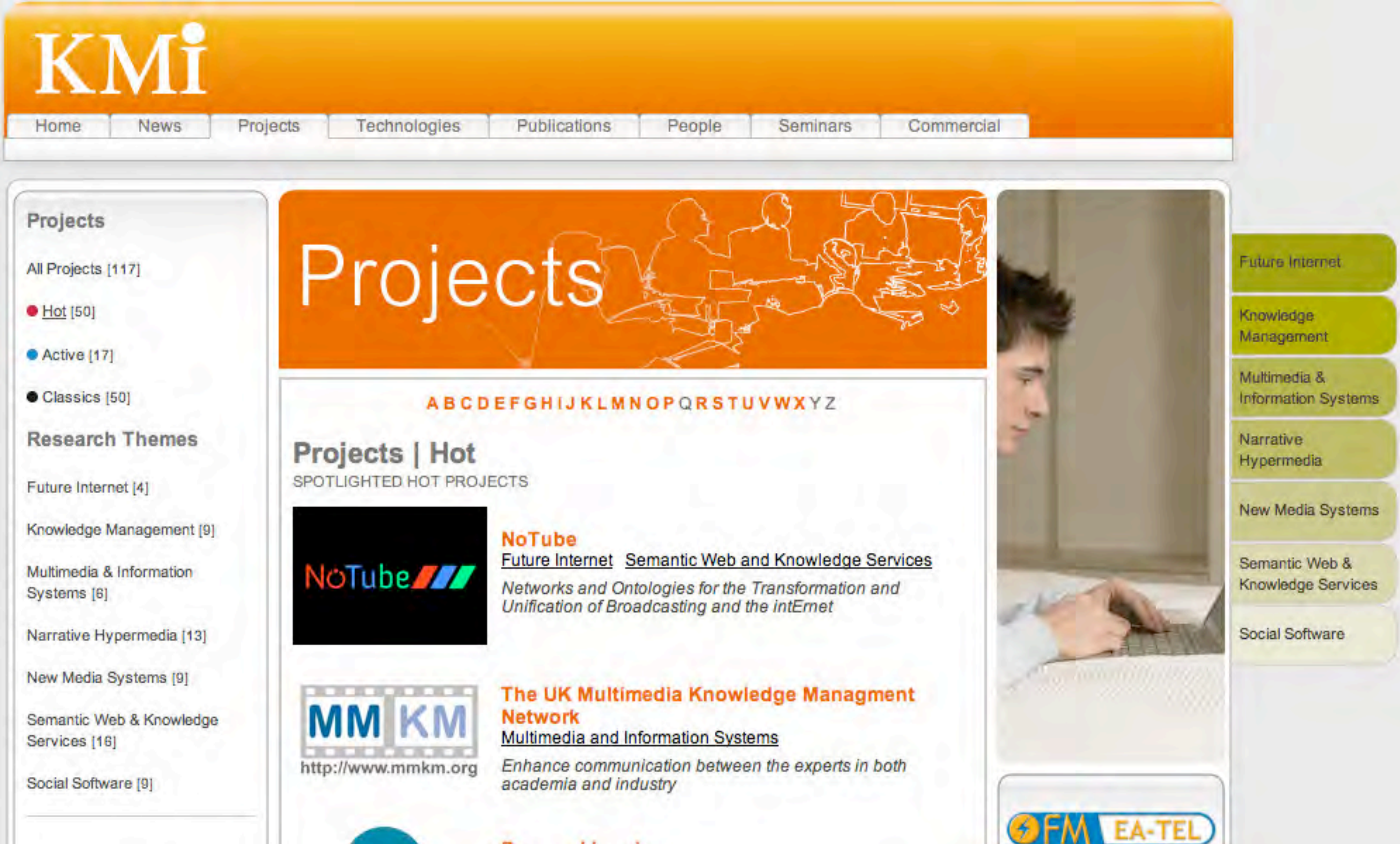
New Media Systems

Semantic Web & Knowledge Services

Social Software

KNOWLEDGE MEDIA KMi INSTITUTE

kmi.open.ac.uk



Since 1995: 117 projects & 67 technologies

Current year

17 live projects

typically £2.5m ext, £1m internal

- 10 EU
- 3 UK
- 1 US
- 3 internal (iTunes U, SocialLearn)

Multimedia information retrieval

1. What is multimedia information retrieval?
2. Metadata and piggyback retrieval
3. Multimedia fingerprinting
4. Automated annotation
5. Content-based retrieval

What is Multimedia?

Within this lecture:

- One or more media
- Possibly interlinked
- Digital
- For communication
  - (not only entertainment)

Multimedia queries

The Twelve Collegia building on Vasilievsky Island in Saint Petersburg is the university's main building and the seat of the rector and administration (the building was constructed on the orders of Peter the Great)

Web-based image searching

"twelve collegia building"

- Google Images
- Bing Images
- Flickr
- Yahoo Images
- ImageToss
- Yandex
- Яндекс

Web-based image searching

Best current practice is a text search:

Find text in filename, anchor text, caption, ...

Text search works by creating a large index:

KNOWLEDGE MEDIA
KMi
INSTITUTE

New search types

doc \ query	text	image	location	speech	sound	humming	motion	Example
text	●							conventional text retrieval
video					●			you roar and get a wildlife documentary
images								
speech	●							type "floods" and get BBC radio news
music						●		hum a tune and get a music piece
sketches								
multimedia								

Exercise

Organise yourself in groups

Discuss with neighbours

- *Two* Examples for different query/doc modes?
- How hard is this? Which techniques are involved?
- *One* example combining different modes

KNOWLEDGE MEDIA KMi INSTITUTE

Exercise

doc \ query	text	image	location	speech	sound	humming	motion
text							
video							
images							
speech							
music							
sketches							
multimedia							

Discuss

- *2* examples
- How hard is it?
- *1* combination

Leaf detection
What are the challenges?

[with Natural History Museum, London, and Goldsmiths]

Venation pattern and shape

Shape is key

[with Frederic Fol Leymarie, Goldsmiths, 2011]

The semantic gap

1m pixels with a spatial colour distribution

faces & vase-like object

Polysemy

1. What is multimedia information retrieval?
2. Metadata and piggyback retrieval
3. Multimedia fingerprinting
4. Automated annotation
5. Content-based retrieval

Metadata

Dublin Core

- simple common denominator: 15 elements such as title, creator, subject, description, ...

METS

- Metadata Encoding and Transmission Standard

MARC 21

- MAchine Readable Cataloguing (harmonised)

MPEG-7

- Multimedia specific metadata standard

MPEG-7

- Moving Picture Experts Group “Multimedia Content Description Interface”
- Not an encoding method like MPEG-1, MPEG-2 or MPEG-4!
- Usually represented in XML format
- Full MPEG-7 description is complex and comprehensive
- Detailed Audiovisual Profile (DAVP)

  [P Schallauer, W Bailer, G Thallinger, “A description infrastructure for audiovisual media processing systems based on MPEG-7”, Journal of Universal Knowledge Management, 2006]

MPEG-7 example

```
<Mpeg7 xsi:schemaLocation="urn:mpeg:mpeg7:schema:2004 ./davp-2005.xsd" ... >
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="AudioVisualType">
   <AudioVisual>
    <StructuralUnit href="urn:x-mpeg-7-pharos:cs:AudioVisualSegmentationCS:root"/>
    <MediaSourceDecomposition criteria="kmi image annotation segment">
     <StillRegion>
      <MediaLocator><MediaUri>http://...392099.jpg</MediaUri></MediaLocator>
      <StructuralUnit href="urn:x-mpeg-7-pharos:cs:SegmentationCS:image"/>
      <TextAnnotation type="urn:x-mpeg-7-pharos:cs:TextAnnotationCS:
        image:keyword:kmi:annotation_1" confidence="0.87">
       <FreeTextAnnotation>tree</FreeTextAnnotation>
      </TextAnnotation>
      <TextAnnotation type="urn:x-mpeg-7-pharos:cs:TextAnnotationCS:
        image:keyword:kmi:annotation_2" confidence="0.72">
       <FreeTextAnnotation>field</FreeTextAnnotation>
      </TextAnnotation>
     </StillRegion>
    </MediaSourceDecomposition>
   </AudioVisual>
  </MultimediaContent>   </Description>  </Mpeg7>
```

KNOWLEDGE MEDIA KMi INSTITUTE

MPEG-7 example

```
<Mpeg7 xsi:schemaLocation="urn:mpeg:mpeg7:schema:2004 ./davp-2005.xsd" ... >
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="AudioVisualType">
   <AudioVisual>
    <StructuralUnit href="urn:x-mpeg-7-pharos:cs:AudioVisualSegmentationCS:root"/>
    <MediaSourceDecomposition criteria="kmi image annotation segment">
     <StillRegion>
      <MediaLocator><MediaUri>http://...392099.jpg</MediaUri></MediaLocator>
      <StructuralUnit href="urn:x-mpeg-7-pharos:cs:SegmentationCS:image"/>
      <TextAnnotation type="urn:x-mpeg-7-pharos:cs:TextAnnotationCS:
        image:keyword:kmi:annotation_1" confidence="0.87">
       <FreeTextAnnotation>tree</FreeTextAnnotation>
      </TextAnnotation>
      <TextAnnotation type="urn:x-mpeg-7-pharos:cs:TextAnnotationCS:
        image:keyword:kmi:annotation_2" confidence="0.72">
       <FreeTextAnnotation>field</FreeTextAnnotation>
      </TextAnnotation>
     </StillRegion>
    </MediaSourceDecomposition>
   </AudioVisual>
  </MultimediaContent>   </Description>  </Mpeg7>
```

Digital libraries

Manage document repositories and their metadata

Greenstone digital library suite

- http://www.greenstone.org/
- interface in 50+ languages (documented in 5)
- knows metadata
- understands multimedia

XML or text retrieval

Piggy-back retrieval

doc \ query	text	image	location	speech	sound	humming	motion
text							
video							
images							
speech							
music							
sketches							
multimedia							

video, images, speech, music, sketches, multimedia } text

KNOWLEDGE MEDIA KMi INSTITUTE

Music to text

0 +7 0 +2 0 -2 0 -2 0 -1 0 -2 0 +2 -4

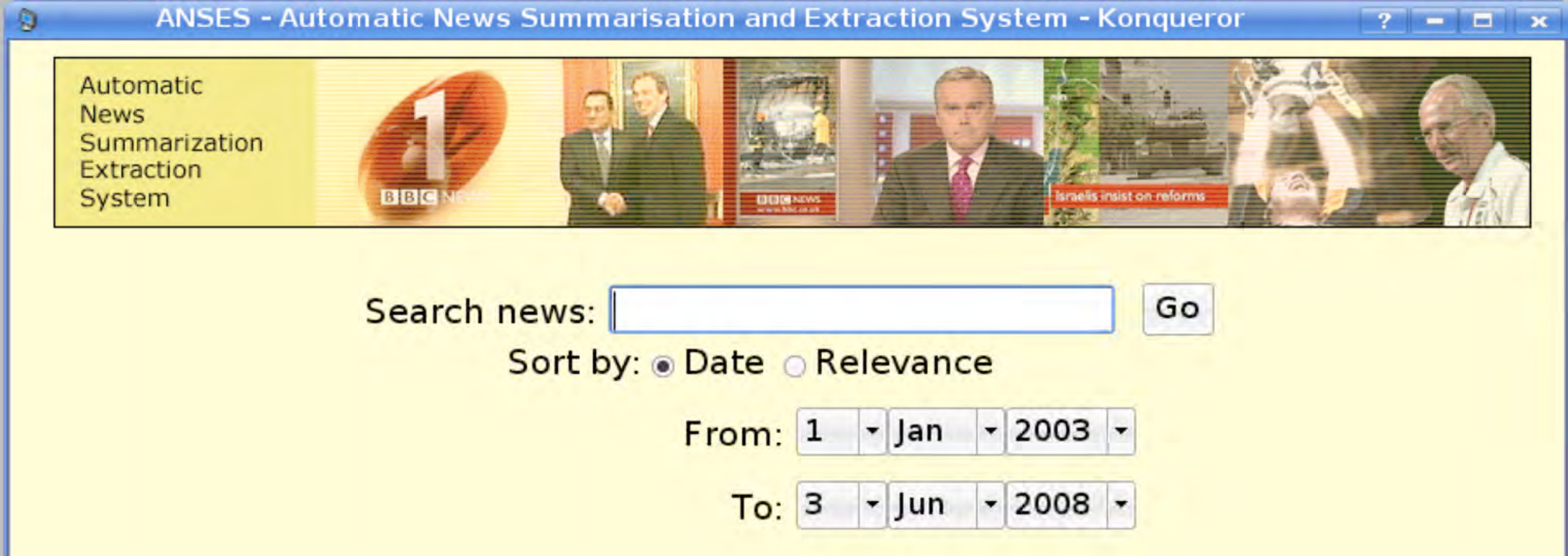
Z G Z B Z b Z b Z a Z b Z B d

ZGZB

GZBZ

ZBZb

[with Doraisamy, J of Intellig Inf Systems 21(1), 2003; Doraisamy PhD thesis 2004]



[technology licensed by Imperial Innovations]

[patent 2004]

[finished PhD: Pickering]

[with Wong and Pickering, CIVR 2003]

[with Lal, DUC 2002]

[Pickering: best UK CS student project 2000 – **national prize**]

ANSES - Automatic News Summarisation and Extraction System - Konqueror

Search news: microsoft Go

Sort by: ◉ Date ○ Relevance

From: 1 Jan 2003

To: 3 Jun 2008

[technology licensed by Imperial Innovations]

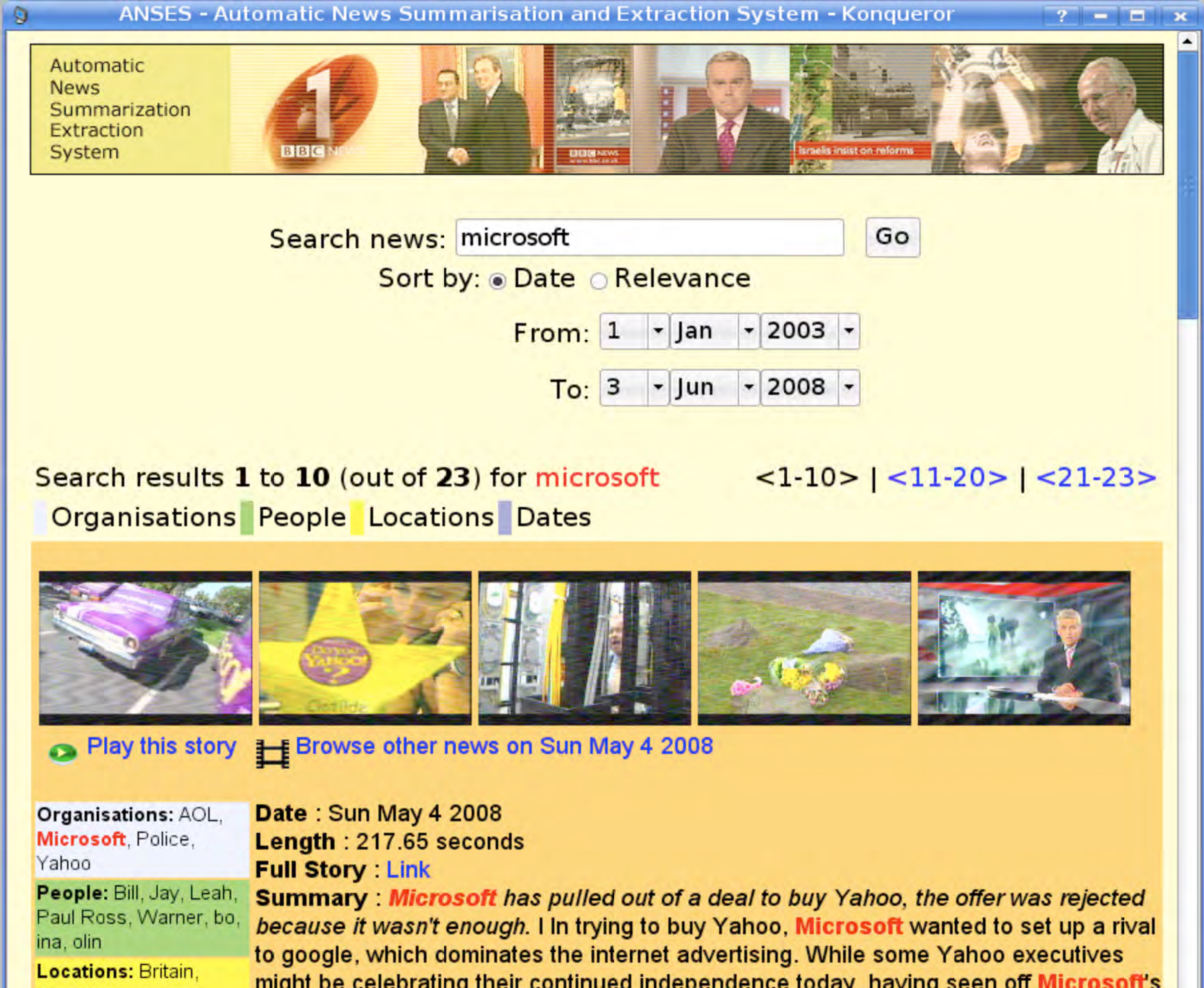
[patent 2004]

[finished PhD: Pickering]

[with Wong and Pickering, CIVR 2003]

[with Lal, DUC 2002]

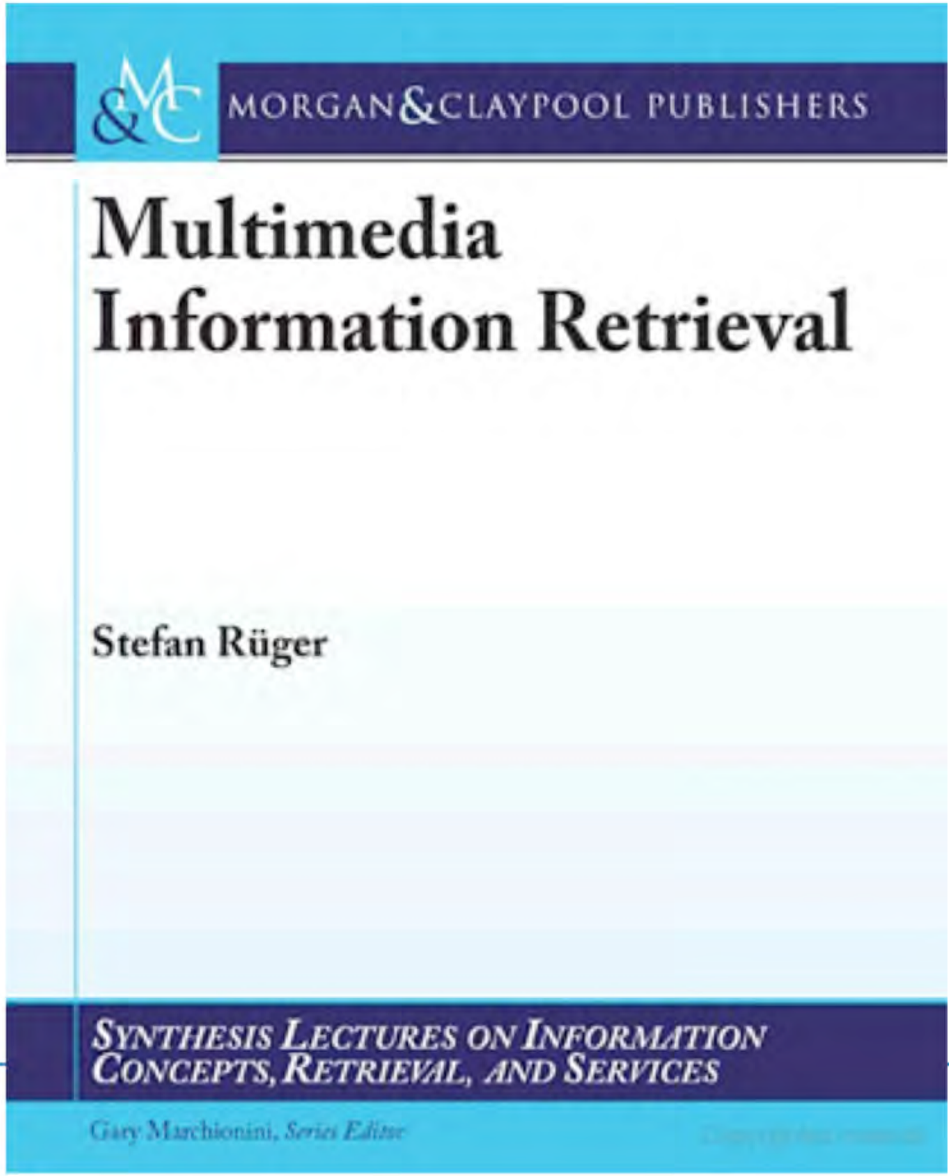
[Pickering: best UK CS student project 2000 – **national prize**]


ANSES - Automatic News Summarisation and Extraction System - Konqueror
Automatic
News
Summarization
Extraction
System
Israelis insist on reforms
Search news: microsoft
Go
Sort by: Date Relevance
From: 1 Jan 2003
To: 3 Jun 2008
Search results 1 to 10 (out of 23) for microsoft
<1-10> | <11-20> | <21-23>
Organisations People Locations Dates
Play this story
Browse other news on Sun May 4 2008
Organisations: AOL, Microsoft, Police, Yahoo
People: Bill, Jay, Leah, Paul Ross, Warner, bo, ina, olin
Locations: Britain,
Date : Sun May 4 2008
Length : 217.65 seconds
Full Story : Link
Summary : Microsoft has pulled out of a deal to buy Yahoo, the offer was rejected because it wasn't enough. I In trying to buy Yahoo, Microsoft wanted to set up a rival to google, which dominates the internet advertising. While some Yahoo executives might be celebrating their continued independence today, having seen off Microsoft's

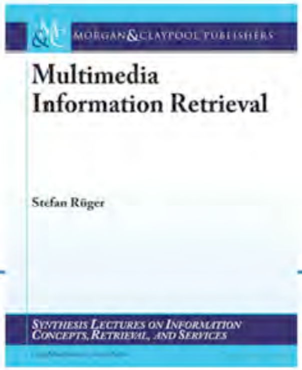

ANSES - Automatic News Summarisation and Extraction System - Konqueror
Search results 1 to 10 (out of 23) for microsoft
<1-10> | <11-20> | <21-23>
Organisations People Locations Dates
Play this story
Browse other news on Sun May 4 2008
Organisations: AOL, Microsoft, Police, Yahoo
People: Bill, Jay, Leah, Paul Ross, Warner, bo, ina, olin
Locations: Britain, Glasgow
Dates: today, tomorrow, yesterday evening
Date : Sun May 4 2008
Length : 217.65 seconds
Full Story : Link
Summary : Microsoft has pulled out of a deal to buy Yahoo, the offer was rejected because it wasn't enough. I In trying to buy Yahoo, Microsoft wanted to set up a rival to google, which dominates the internet advertising. While some Yahoo executives might be celebrating their continued independence today, having seen off Microsoft's unwanted attentions, they might already been dreading stock markets pening tomorrow. Both Microsoft and Yahoo have come a long way since being ormed in garages, both sets have earned billions along the way. Alternative Leah yahoo may look merge with AOL, owned by Time mre whAO own by ime Warner, but it would have to fast, because AOL might also be under Microsoft's radar.
Play this story
Browse other news on Tue Sep 25 2007

KNOWLEDGE MEDIA KMi INSTITUTE

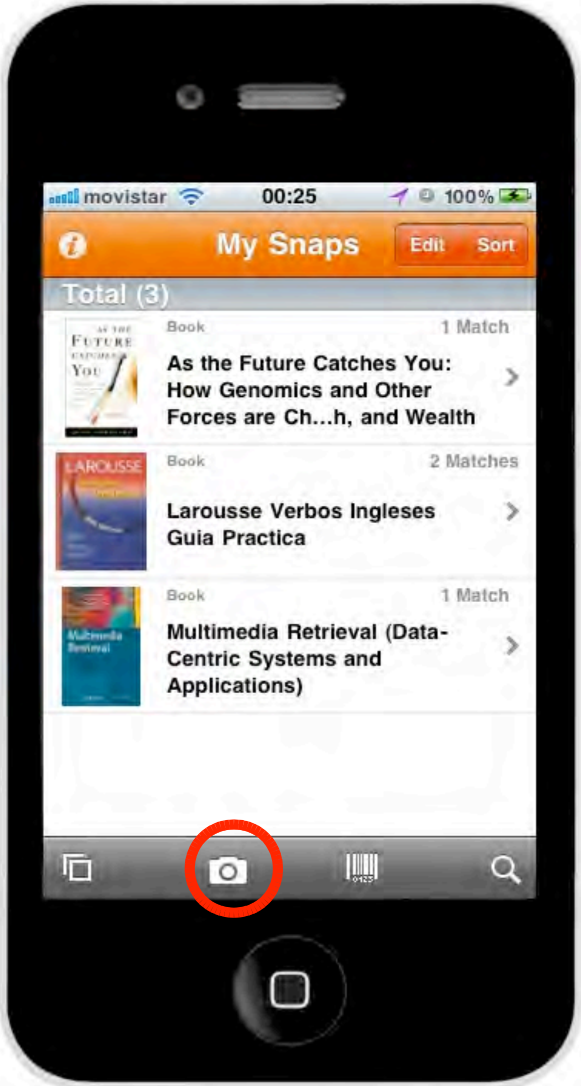
Near-duplictate detection: Cool access mode!



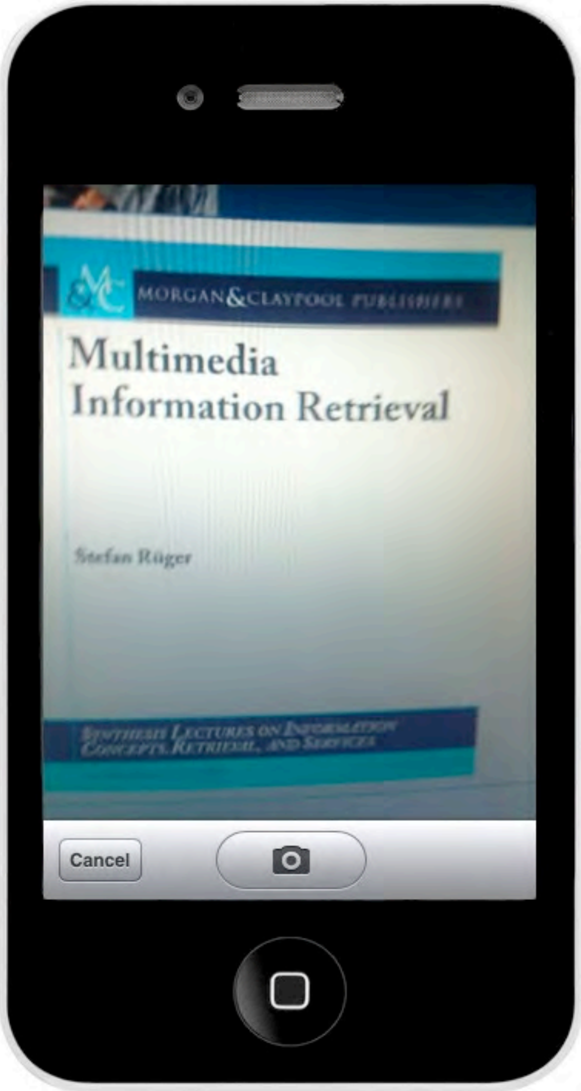
Snaptell: Book, CD and DVD covers



Snaptell: Book, CD and DVD covers



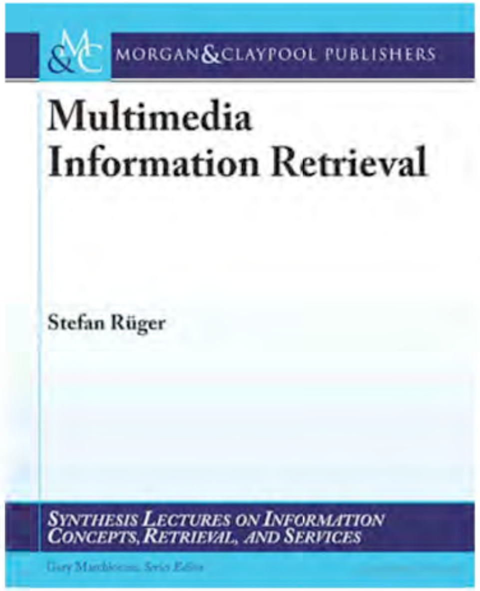
Snaptell: Book, CD and DVD covers



Snaptell: Book, CD and DVD covers

Snaptell: Book, CD and DVD covers

Link from real world to databases



doi: 10.2200/S00244ED1V01Y200912ICR010

KNOWLEDGE MEDIA INSTITUTE

The Open University's Spot & Search

Scott Forrest: E=MC squared

"Between finished surface texture and raw quarried stone. Between hard materials and soft concepts. Between text and context."

More information

[with Suzanne Little]

KNOWLEDGE MEDIA KMi INSTITUTE

Spot & Search

[with Suzanne Little]

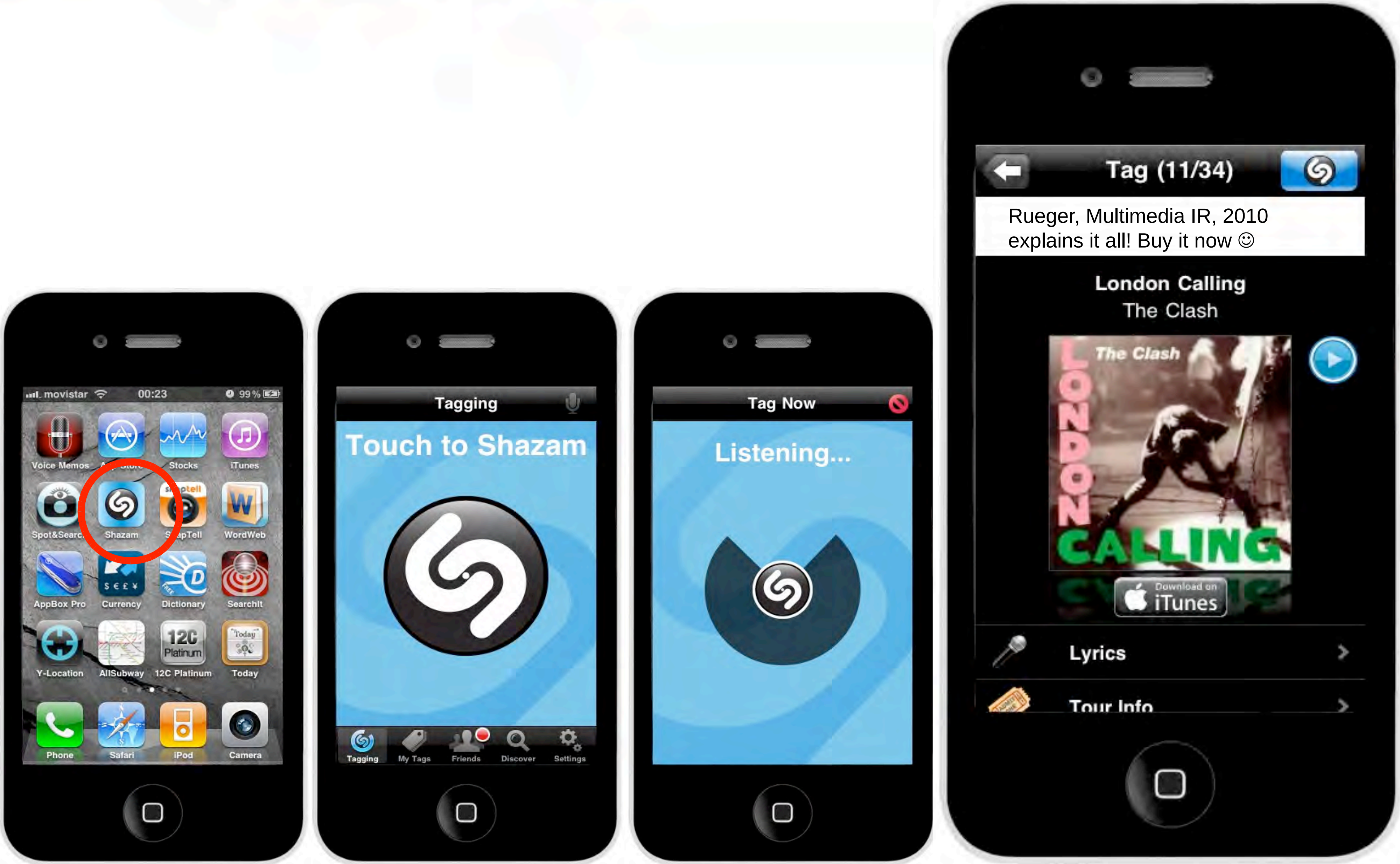
Near duplicate detection

Works well in 2d: CD covers, wine labels, signs, ...

Less so in near 2d: buildings, vases, ...

Not so well in 3d: faces, complex objects, ...

Shazam


movistar
00:23
99 %
Voice Memos
Stocks
iTunes
Spot&Search
Shazam
WordWeb
AppBox Pro
Currency
Dictionary
SearchIt
Y-Location
AllSubway
12C Platinum
Today
Phone
Safari
iPod
Camera
Tagging
Touch to Shazam
Tagging
My Tags
Friends
Discover
Settings
Tag Now
Listening...
Tag (11/34)
Rueger, Multimedia IR, 2010
explains it all! Buy it now ☺
London Calling
The Clash
The Clash
LONDON
CALLING
Download on
iTunes
Lyrics
Tour Info

Near duplicate detection
Exercise

Find applications for near-duplicate detection

- be imaginative: the more “outragous” the better
- can be other media types (audio, smells, haptic, ...)
- can be hard to do

Near-duplicate detection
Where are the challenges?

[Victoria and Albert museum, London, ceramics collection, 2010]

How does near-duplicate detection work?

Fingerprinting technique

1 Compute salient points

2 Extract “characteristics” from vincinity (feature)

3 Make invariant under rotation & scaling

4 Quantise: create visterms

5 Index as in text search engines

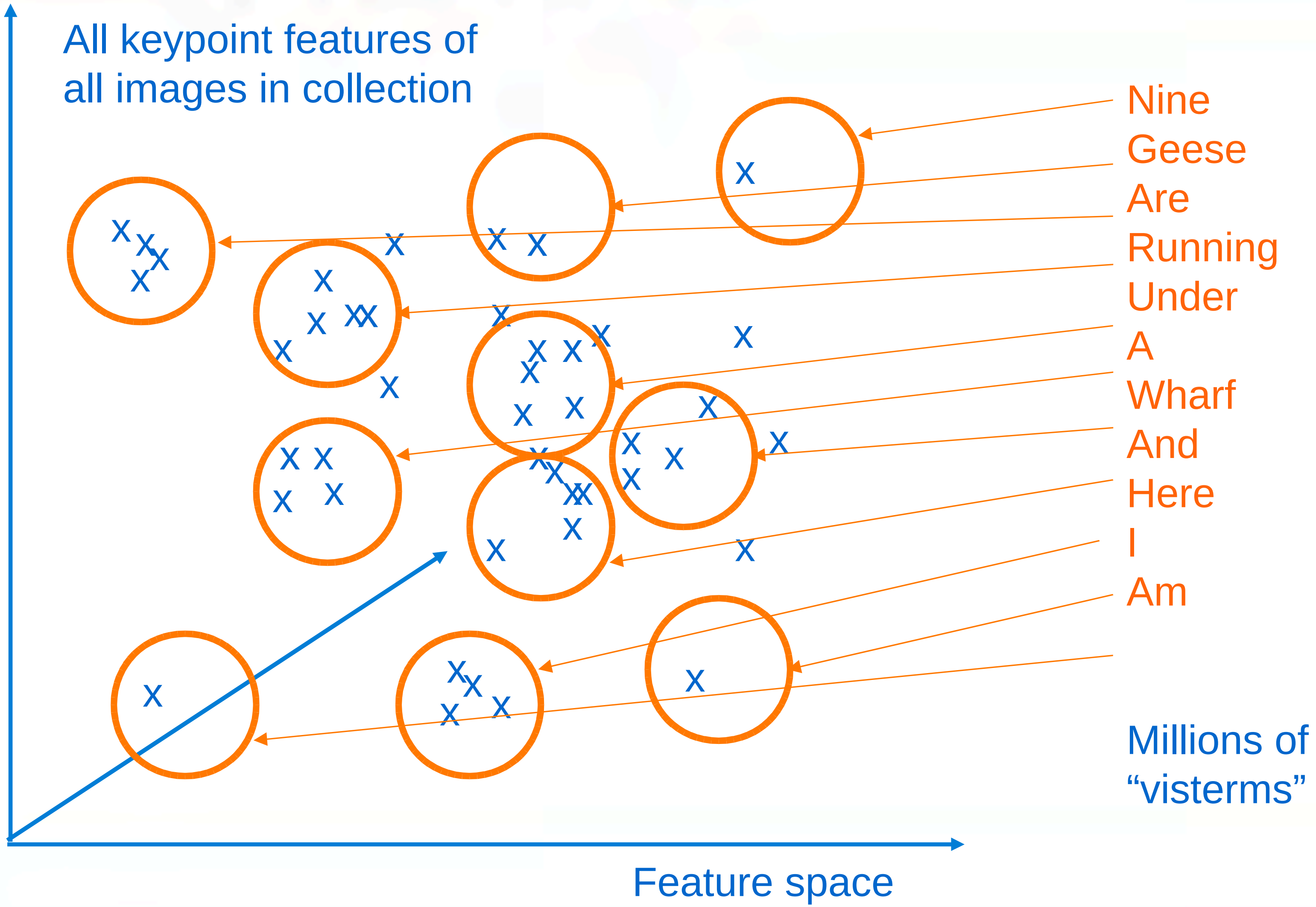
6 Check/enforce spatial constraints after retrieval

NDD: Compute salient points and features

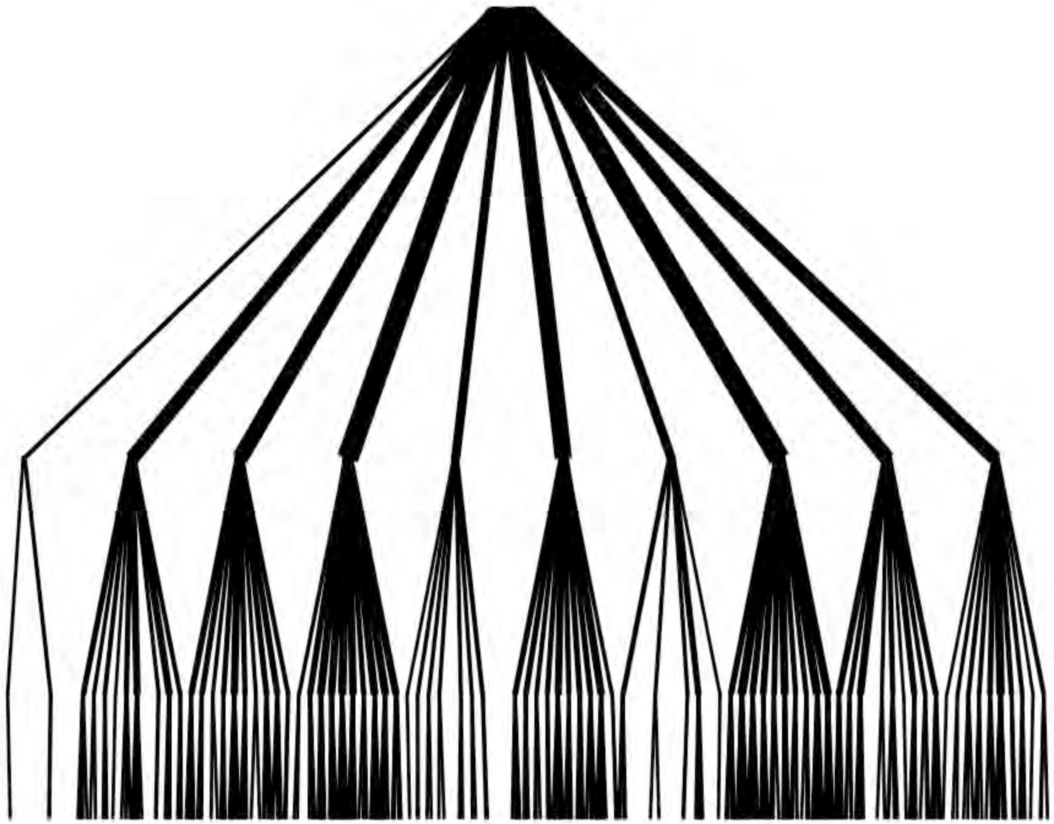
Eg, SIFT features: each salient point described by a feature vector of 128 numbers; the vector is invariant to scaling and rotation

[Lowe2004 – http://www.cs.ubc.ca/~lowe/keypoints/]

NDD: Keypoint feature space clustering



Clustering
Hierarchical k-means



[Nister and Stewenius, CVPR 2006]

NDD: Encode all images with visterms

Jkjh Geese Bjlkj Wharf
Ojkkjhhj Kssn Klkekjl Here
Lkjkll Wjjkll Kkjlk Bnm
Kllkgjg Lwoe Boerm …

NDD: query like text

At query time compute salient points, keypoint features and visterms

Query against database of images represented as bag of vistems

Query

socialllearn →

Joiu Gddwd Bipoi Wueft
Oiooiuui Kwwn Kpodoip Hdfd
Loiopp Wiiopp Koipo Bnm
Kppoyiy Lsld Bldfm …

[with Suzanne Little]

NDD: Check spatial constraints

[with Suzanne Little, SocialLearn project]

KNOWLEDGE MEDIA KMi INSTITUTE

How does near-duplicate detection work?

Fingerprinting technique

1 Compute salient points

2 Extract “characteristics” from vincinity (feature)

3 Make invariant under rotation & scaling

4 Quantise: create visterms

5 Index as in text search engines

6 Check/enforce spatial constraints after retrieval

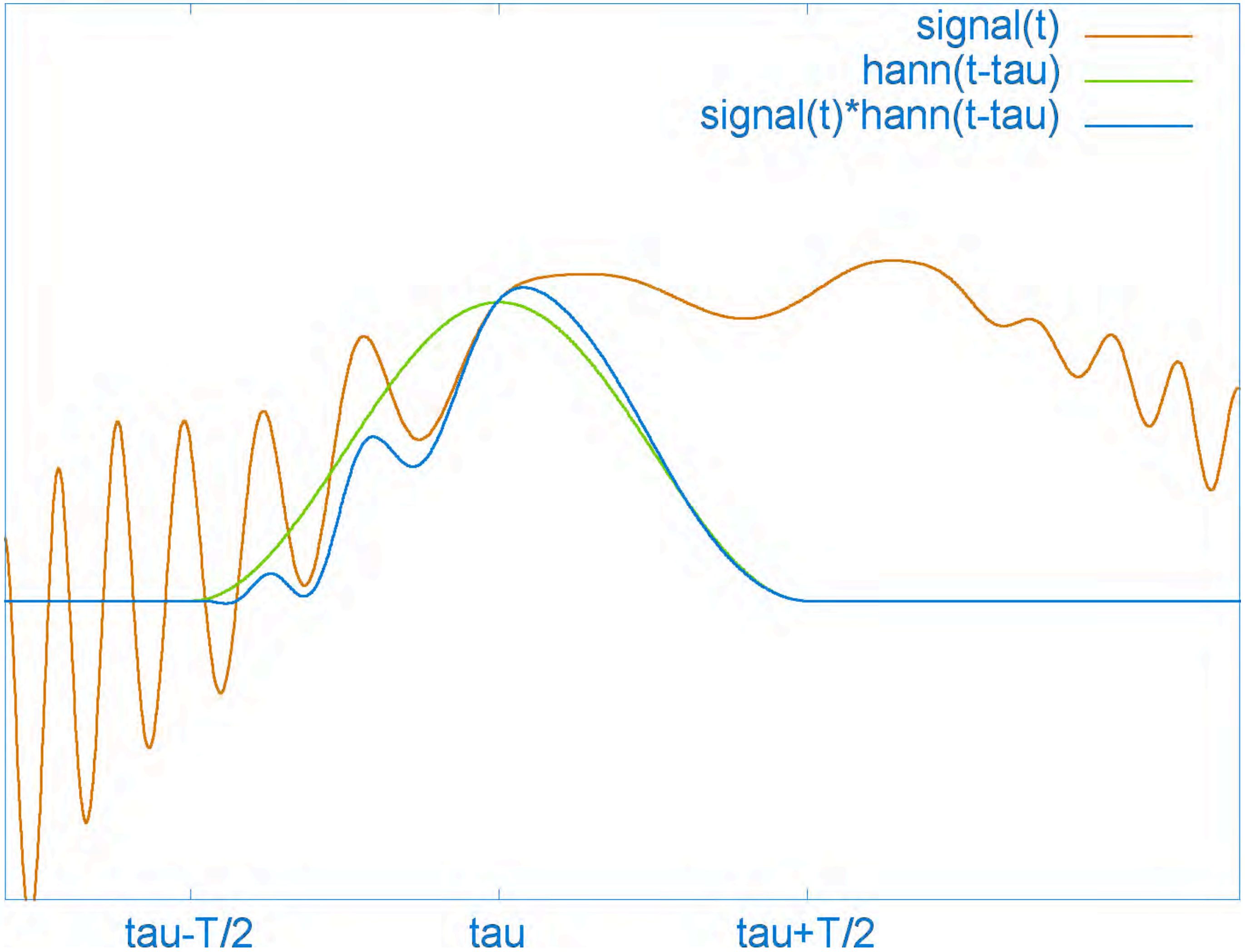
How Shazam works - Spectrogram

Compute energy for all (frequency,time) pairs using a Fourier transform under a Hann window w

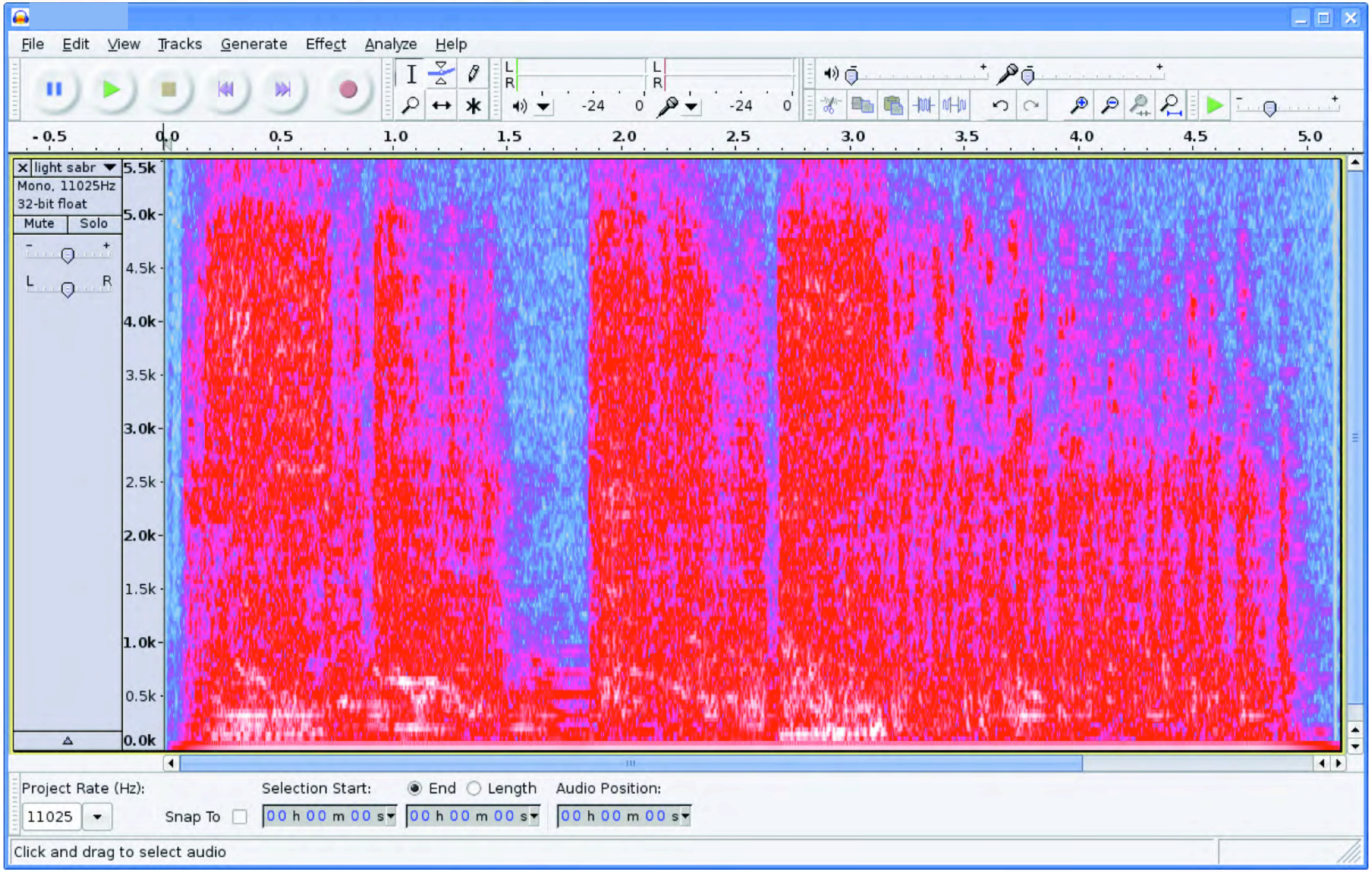
$$\text{spectrogram}(f,\tau) = \left| \int_{-\infty}^{\infty} s(t)w(t-\tau)e^{ift}dt \right|^2$$

$$w(t) = \begin{cases} \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi t}{T}\right) & \text{if } t \in [-T/2, T/2] \\ 0 & \text{otherwise} \end{cases}$$

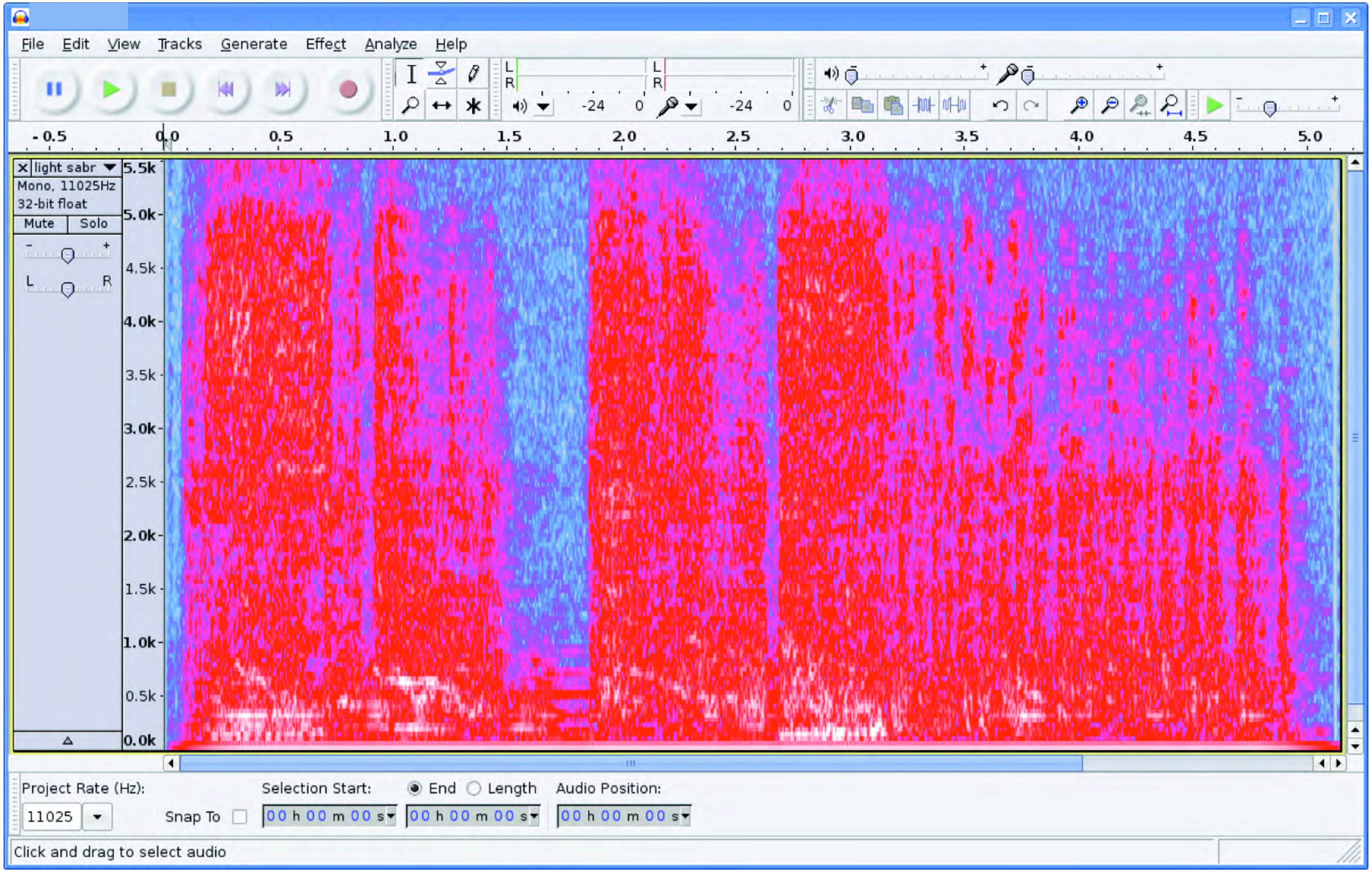
Hann window application



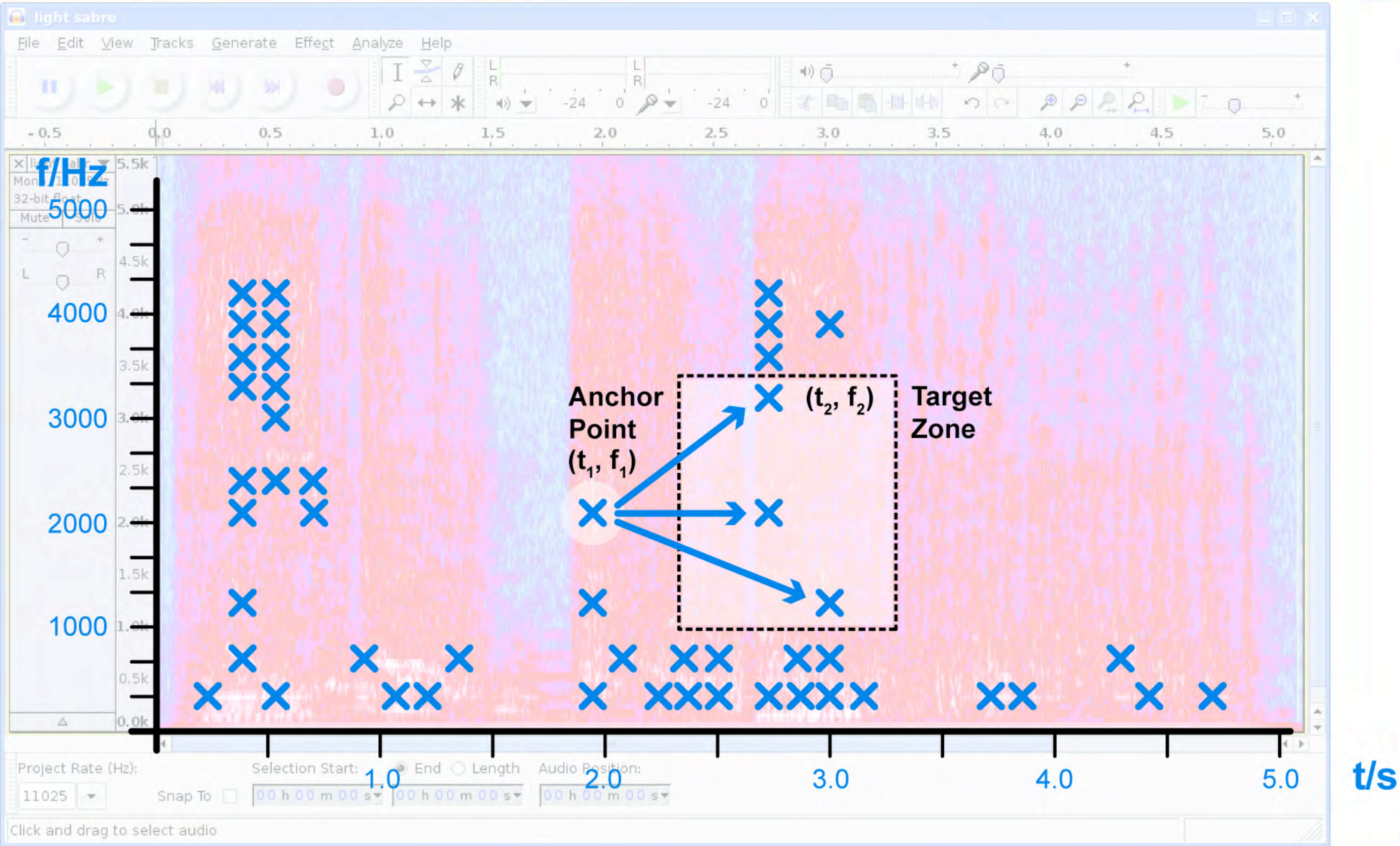
How Shazam works: audio fingerprinting



How Shazam works: audio fingerprinting



Salient points



Encoding: $(f_1, f_2, t_2\text{-}t_1)$ hashes to (t_1, id)

[Wang(2003), An industrial-strength search algotithm, ISMIR]

Temporal consistency check of query

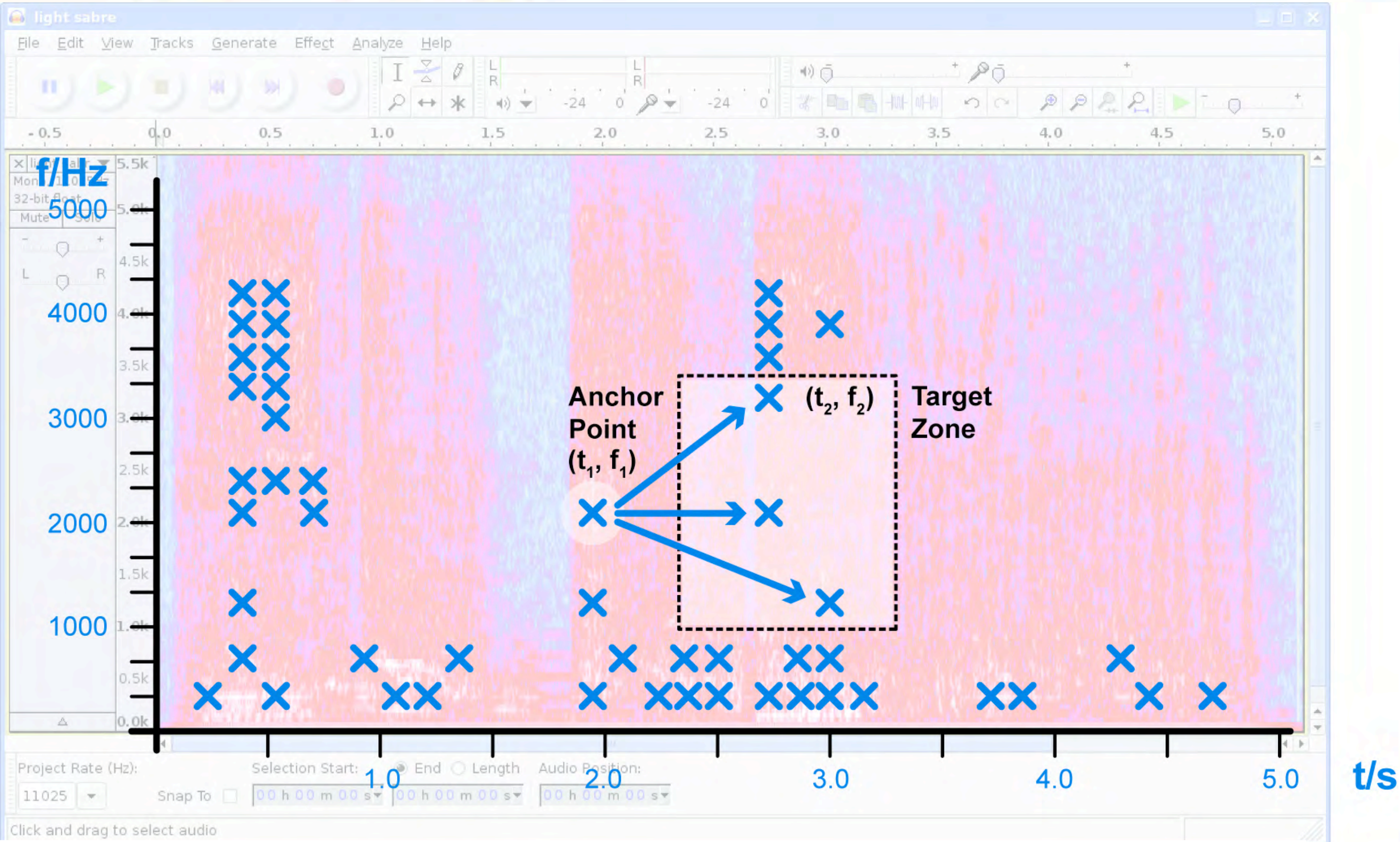
Every query vector (f_1,f_2, t^q_2-t^q_1) is matched to the database.

You get a list of possible (t^{id}_1, id) values (some are false positives).

Create a histogram of t^{id}_1-t^q_1 values (temporal consistency check!)

A substantial peak in this histogram means that the query has matched song id at time offset t^{id}_1-t^q_1.

Entropy considerations



Specificity: Encoding $(f_1, f_2, t_2\text{-}t_1)$ to use 30 bit

Exercise
Shazam's constellation pairs

Assume that the typical survival probability of each 30-bit constellation pair after deformations that we still want to recognise is p, and that this process is independent per pair. Which encoding density, ie, the number of constellation pairs per second, would you need on average so that a typical query of 10 seconds exhibits at least 10 matches in the right song with a probability of at least 99.99%?

Under these assumptions, further assuming that the constellation pair extraction looks like a random independent and identically distributed number, what is the false positive rate for a database of 4 million songs each of which is 5 minutes long on average?

Exercise
Shazam's constellation pairs

Which encoding density would you need on average so that a typical query of 10 seconds exhibits at least 10 matches in the right song with a probability of at least 99.99%?

- approximately 1 match per second needed (n = pairs/second):

$$1 - (1 - p)^n = 0.9999$$

$$n = \log(0.0001) / \log(1 - p)$$

Exercise
Shazam's constellation pairs

Which encoding density would you need on average so that a typical query of 10 seconds exhibits at least 10 matches in the right song with a probability of at least 99.99%?

- Exact solution: binomial distribution

$$n = \min\left\{ m \;\middle|\; \sum_{i \geq 10}^{10m} \binom{10m}{i} (1-p)^{10m-i} p^i \geq 0.9999 \right\}$$

Exercise
Shazam's constellation pairs

Which encoding density would you need on average so that a typical query of 10 seconds exhibits at least 10 matches in the right song with a probability of at least 99.99%?

- Large n: approximate binomial distribution with N(np, sqrt(np(1-p)))

Exercise
Shazam's constellation pairs

Assuming that the constellation pair extraction looks like a random independent and identically distributed number, what is the false positive rate for a database of 4 million songs each of which is 5 minutes long on average?

Zero:

5min = 30*10sec
(assume distinctive 2^{30})
$m = 2^{-30}$

p(query matches one segment) approx m^{10} approx 2^{-300}
$1-(1-p(qms))^{(30*4e6)}$ approx $120e6*m^{10}$
still near zero

Philips Research

Divide frequency scale into 33 frequency bands between 300 Hz and 2000 Hz
Logarithmic spread – each frequency step is 1/12 octave, ie, one semitone

Divide time axis into blocks of 256 windows of 11.6 ms (3 seconds)

E(m,n) is the energy of the m-th frequency at n-th time in spectrogram

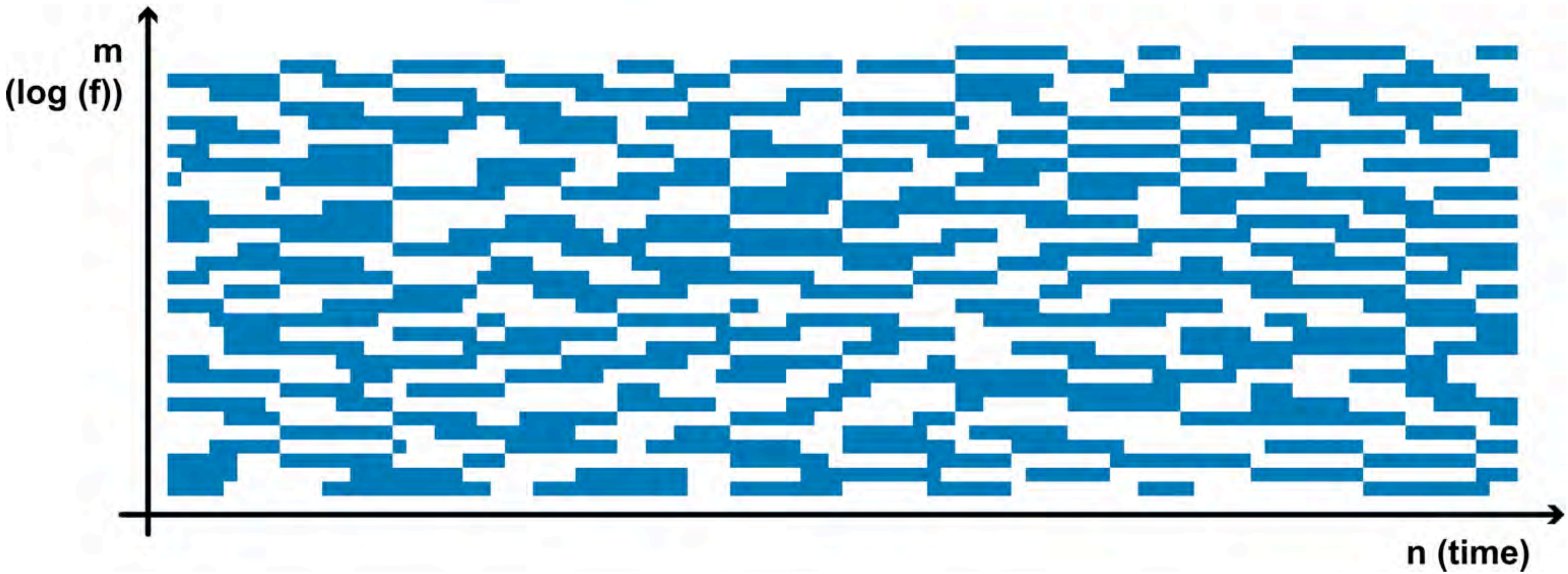
For each block extract 256 sub-fingerprints of 32 bits each

$$b(m, n) = \text{sign}\left([E(m, n) - E(m+1, n)] - [E(m, n+1) - E(m+1, n+1)]\right)$$

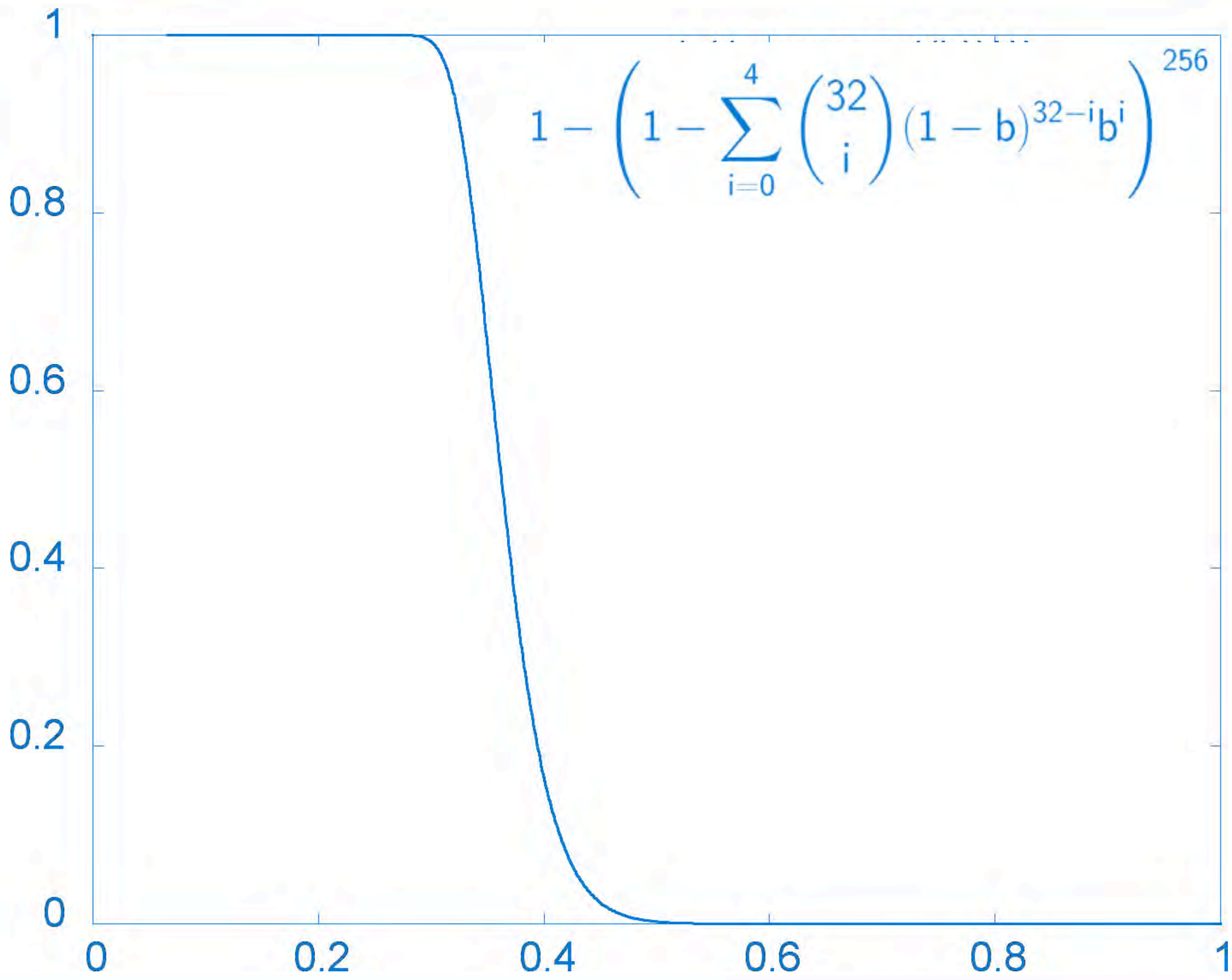
$$0 \leq m \leq 31 \text{ (frequency)}$$
$$0 \leq n \leq 255 \text{ (time)}$$

[Haitsma and Kalker, 2003]

Partial fingerprint block



Probability of at least one sub-fingerprint surviving with no more than 4 errors



Quantisation through locality sensitive hashing (LSH)

$$h^i : \mathbb{R}^d \longrightarrow \mathbb{Z}$$
$$v \longmapsto h^i(v) = \left\lfloor \frac{a^i v + b^i}{w} \right\rfloor$$

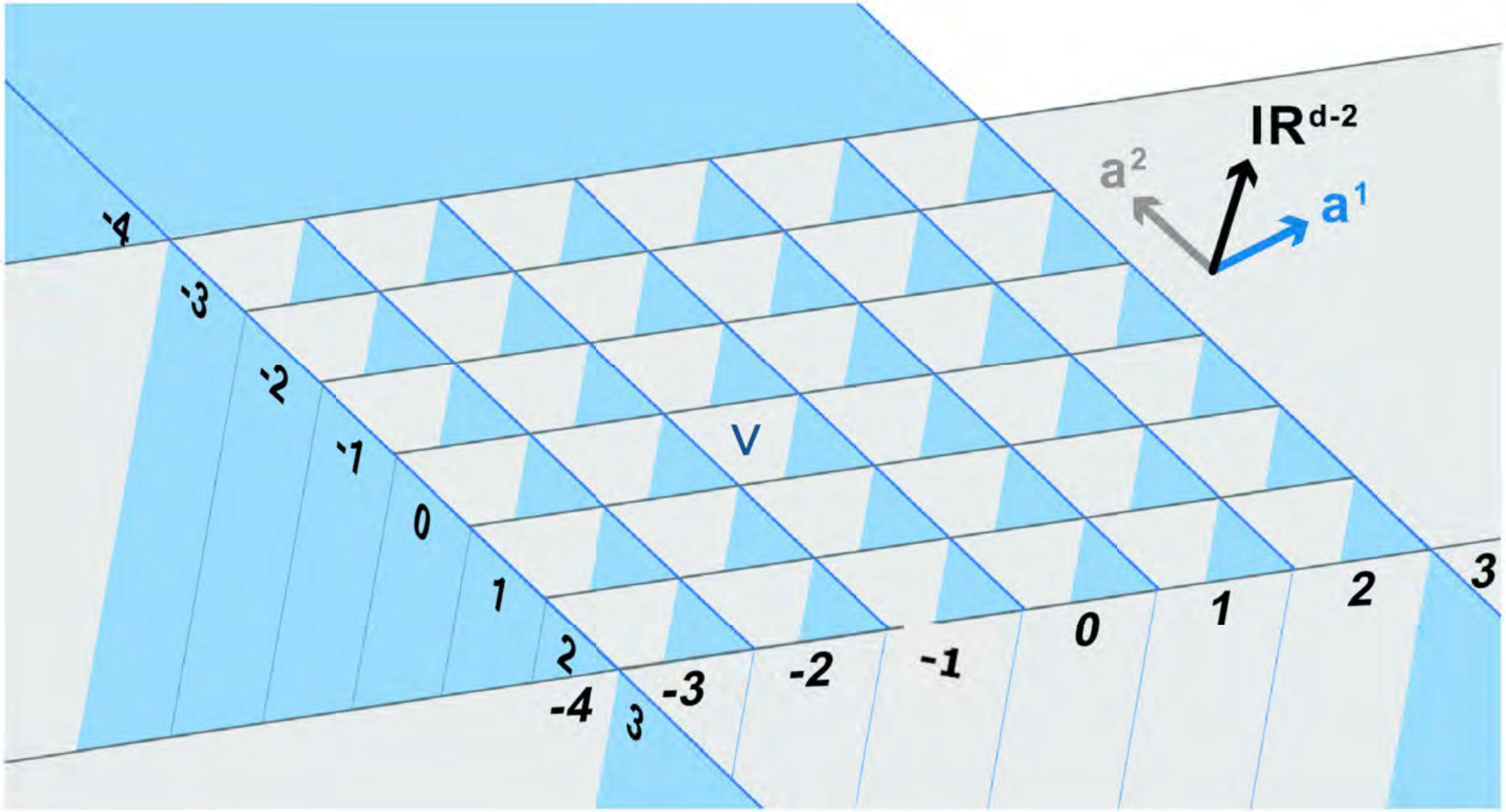
$a^i \in \mathbb{R}^d$ is a random Gaussian-distributed vector

$w \in \mathbb{R}^+$ is a constant

$b^i \in [0, w)$ is a random number

$h(v) = (h^1(v), h(^2(v), \ldots, h^k(v))$ is the LSH hash vector.

Quantisation LSH hashes



Vector v: h(v) = (-1, 0)

Redundancy is key

Use L independent hash vectors of k components each both for the query and for each multimedia object.

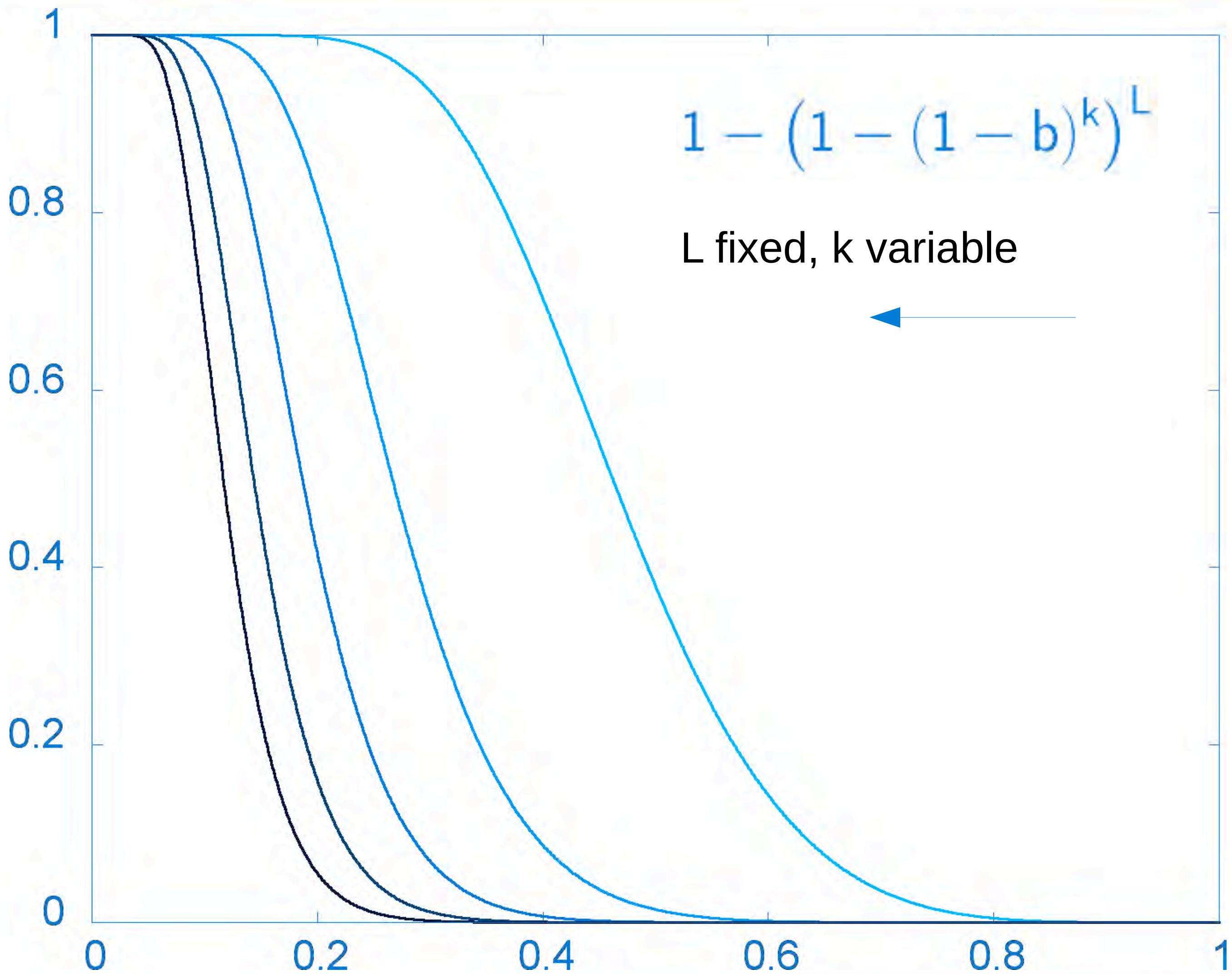
Database elements that match at least m out of L times are candidates for nearest neighbours.

Chose w, k and L (wisely) at runtime

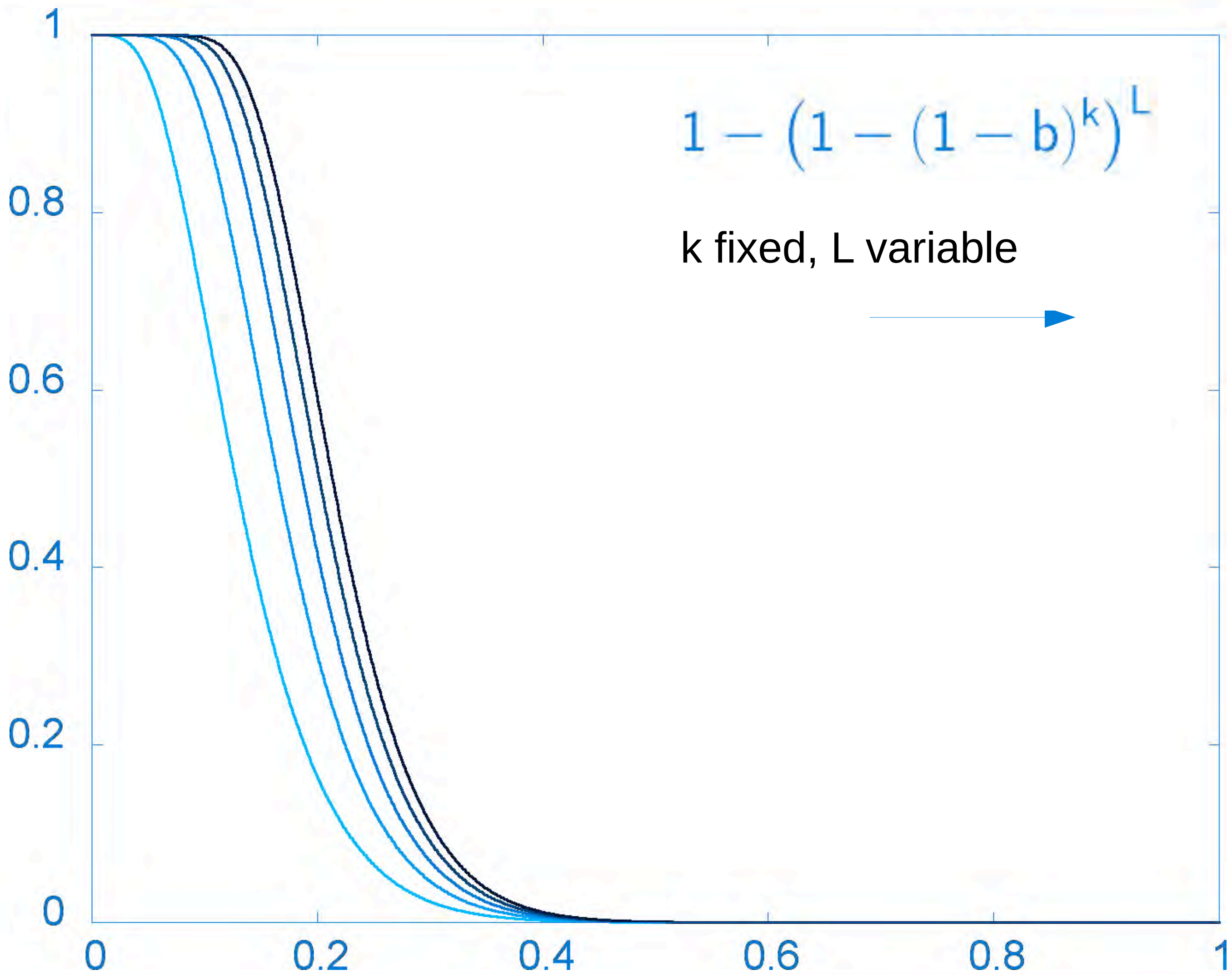
- w determines granularity of bins, ie, # of bits for $h^i(v)$

- k and L determine probability of matching

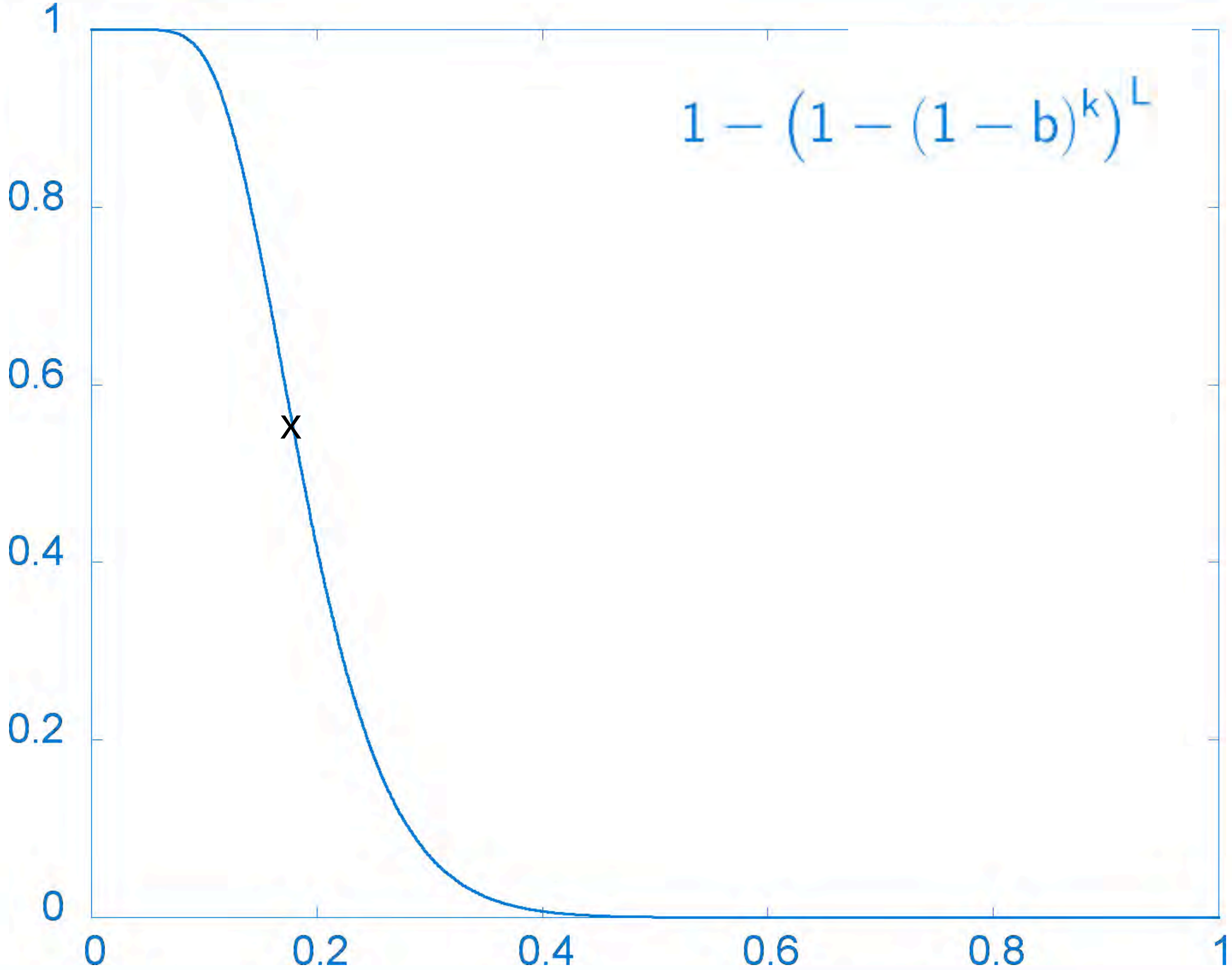
Prob(min 1 match out of L)



Prob(min 1 match out of L)



Exercise: compute inflection point



Min hash
Estimate discrete set overlap

$$\mathrm{sim}(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

An example
4 documents

D_1 = Humpty Dumpty sat on a wall,

D_2 = Humpty Dumpty had a great fall.

D_3 = All the King's horses, And all the King's men

D_4 = Couldn't put Humpty together again!

Surrogate docs after stop word removal and stemming

A_1 = {humpty, dumpty, sat, wall}
A_2 = {humpty, dumpty, great, fall}
A_3 = {all, king, horse, men}
A_4 = {put, humpty, together, again}

Equivalent term-document matrix

	A_1	A_2	A_3	A_4
humpty	1	1	0	1
dumpty	1	1	0	0
sat	1	0	0	0
wall	1	0	0	0
great	0	1	0	0
fall	0	1	0	0
all	0	0	1	0
king	0	0	1	0
horse	0	0	1	0
men	0	0	1	0
put	0	0	0	1
together	0	0	0	1
again	0	0	0	1

Similarity between two docs

$$\text{sim}(A_i, A_j) = \frac{c_{11}}{c_{11}+c_{10}+c_{01}}$$

$$c_{xy} = \text{number of (x,y) rows}$$

Important observation
c_{00} is unused!

	A_1	A_2	A_3	A_4
humpty	1	1	0	1
dumpty	1	1	0	0
sat	1	0	0	0
wall	1	0	0	0
great	0	1	0	0
fall	0	1	0	0
all	0	0	1	0
king	0	0	1	0
horse	0	0	1	0
men	0	0	1	0
put	0	0	0	1
together	0	0	0	1
again	0	0	0	1

Estimation of similarity through random permutations

π_1 = (dumpty, men, again, put, great, humpty, wall, horse, king, sat, fall, together, all)

π_2 = (fall, put, all, again, dumpty, sat, men, great, wall, king, horse, humpty, together)

π_3 = (horse, dumpty, wall, humpty, great, again, sat, all, men, together, put, king, fall)

π_4 = (king, humpty, men, together, great, fall, horse, all, dumpty, wall, sat, again, put)

Surrogate documents form random permutations

Keep first occurring word of A_i in π_j for dense surrogate representation

	A_1	A_2	A_3	A_4
π_1	dumpty	dumpty	men	again
π_2	dumpty	fall	all	put
π_3	dumpty	dumpty	horse	humpty
π_4	humpty	humpty	king	humpty

Surrogate documents form random permutations

	A_1	A_2	A_3	A_4
π_1	dumpty	dumpty	men	again
π_2	dumpty	fall	all	put
π_3	dumpty	dumpty	horse	humpty
π_4	humpty	humpty	king	humpty

Estimate sim(A_2,A_4) = 1/4
(proportion of co-inciding words)

SIFT

Scale Invariant Feature Transform

"distinctive invariant image features that can be used to perform reliable matching between different views of an object or scene."

Invariant to image scale and rotation.

Robust to substantial range of affine distortion, changes in 3D viewpoint, addition of noise and change in illumination.

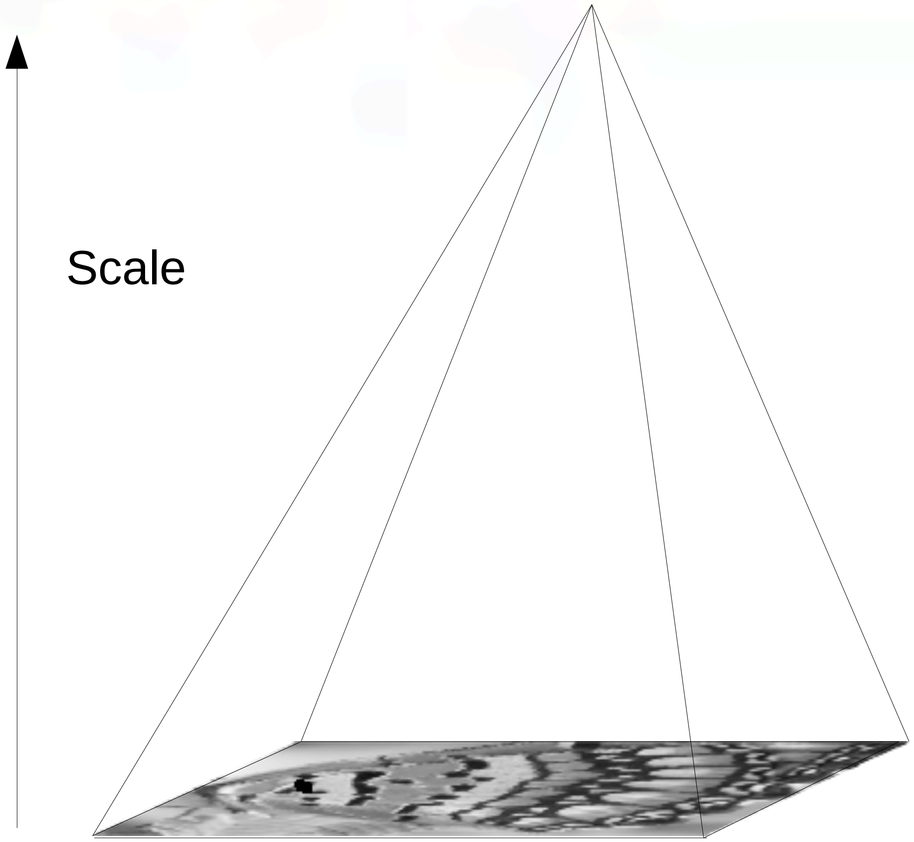
[Lowe, D.G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60, 2, pp. 91-110.]

SIFT Implementation

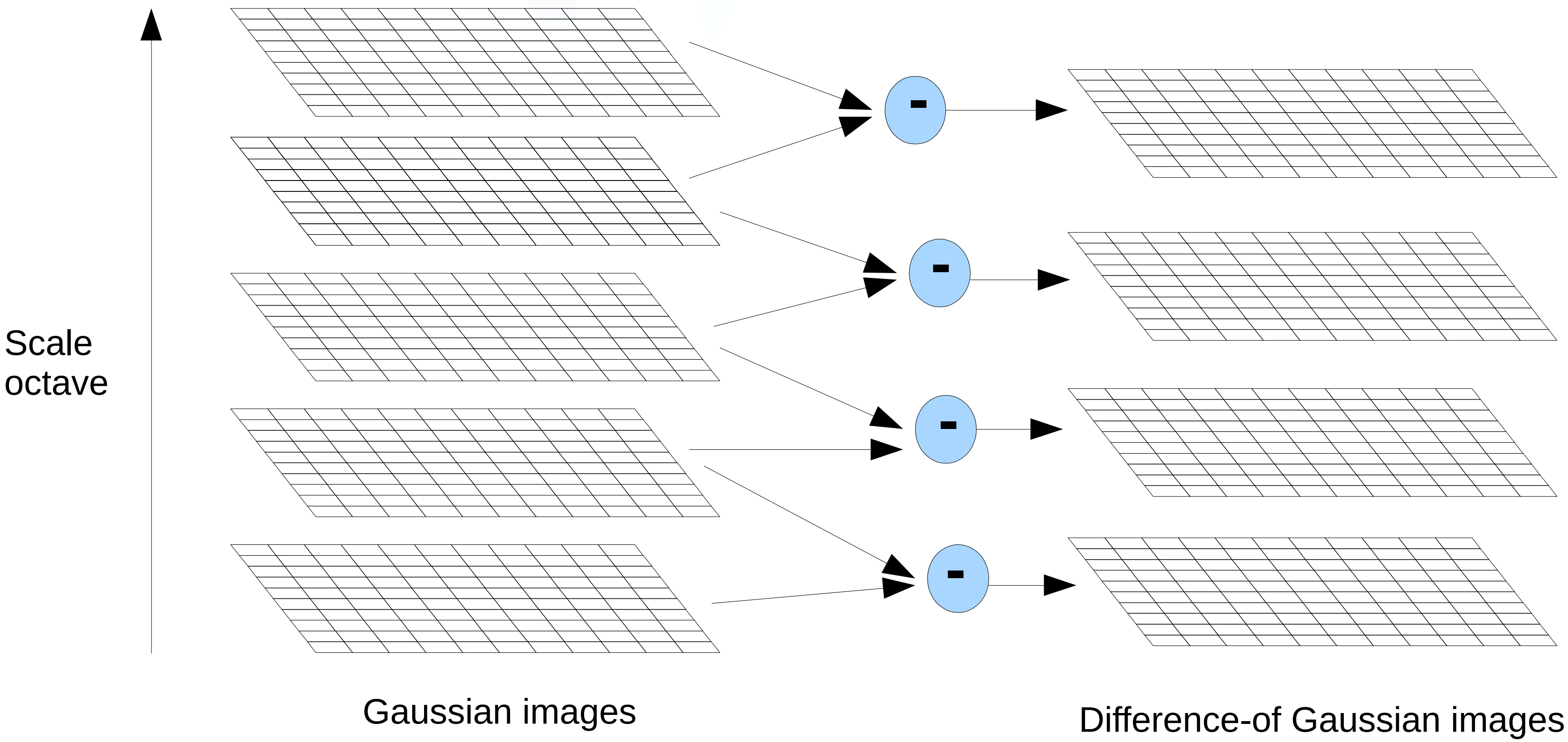
For a given image:

- Detect scale space extrema
- Localise candidate keypoints
- Assign an orientation to each keypoint
- Produce keypoint descriptor

A scale space visualisation



Difference of Gaussian image creation



Gaussian blur illustration

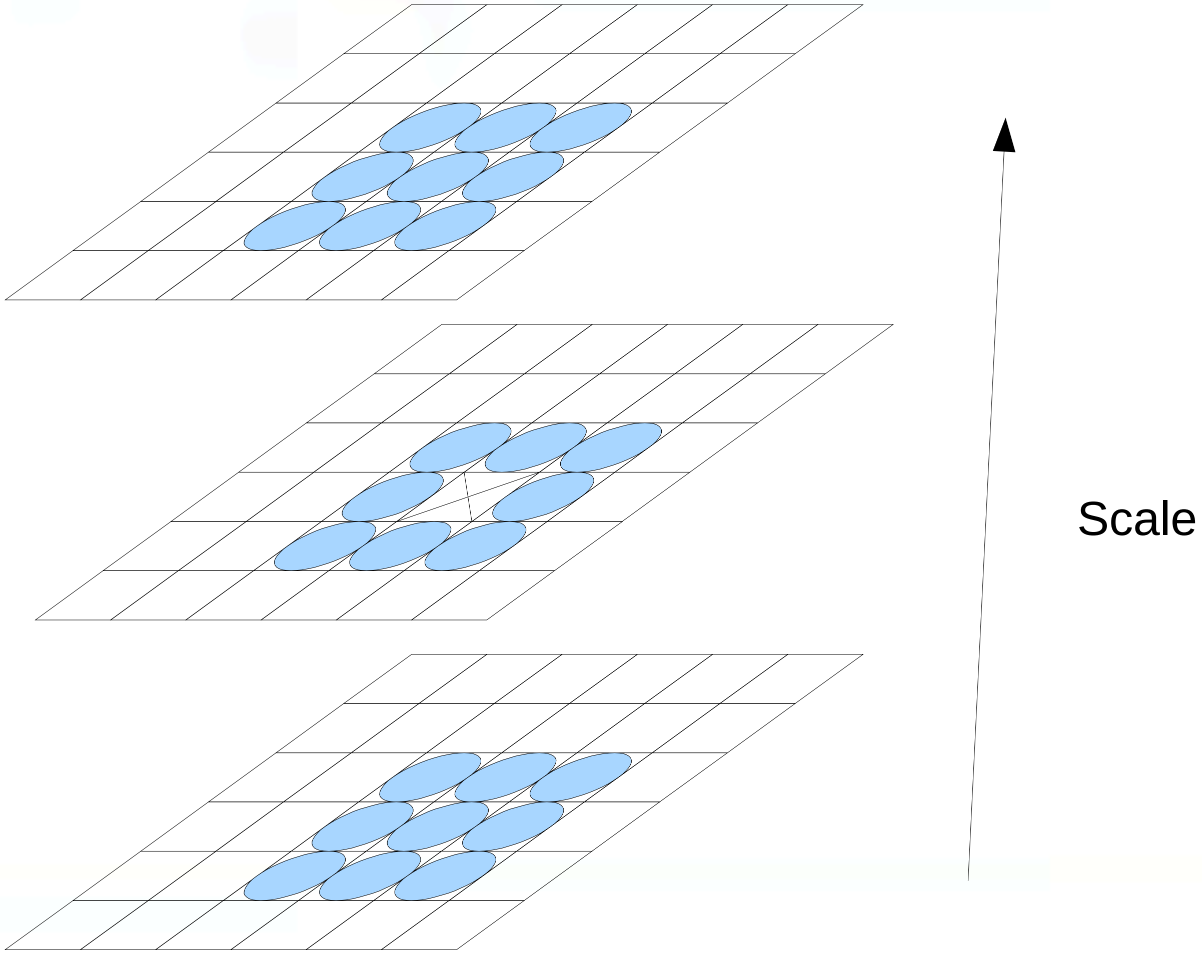
Difference of Gaussian illustration

The SIFT keypoint system

Once the Difference of Gaussian images have been generated:

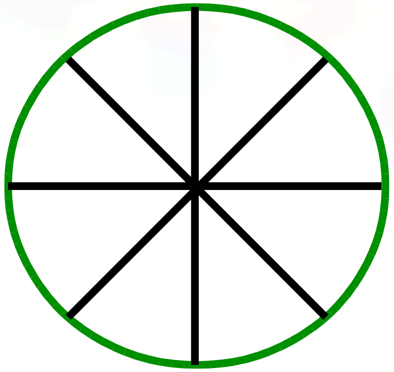
- Each pixel in the images is compared to 8 neighbours at same scale.
- Also compared to 9 corresponding neighbours in scale above and 9 corresponding neighbours in the scale below.
- Each pixel is compared to 26 neighbouring pixels in 3x3 regions across scales, as it is not compared to itself at the current scale.
- A pixel is selected as a SIFT keypoint only either if its intensity value is extreme.

Pixel neighbourhood comparison

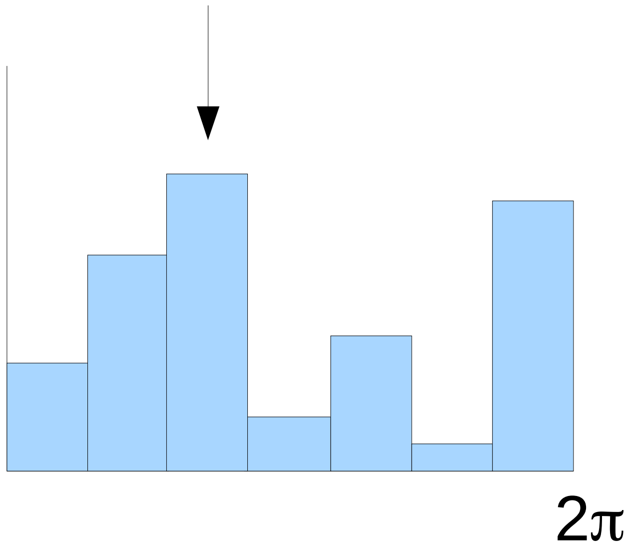


KNOWLEDGE MEDIA KMi INSTITUTE

Orientation assignment

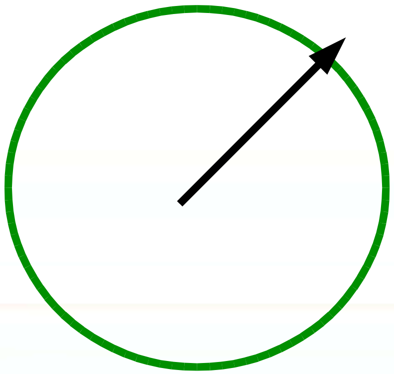


Orientation histogram with 36 bins – one per 10 degrees.



Each sample weighted by gradient magnitude and Gaussian window.

Canonical orientation at peak of Smoothed histogram.



Where two or more orientations are detected, keypoints created for each orientation.

The SIFT keypoint descriptor

We now have location, scale and orientation for each SIFT keypoint ("keypoint frame").

→ descriptor for local image region is required.

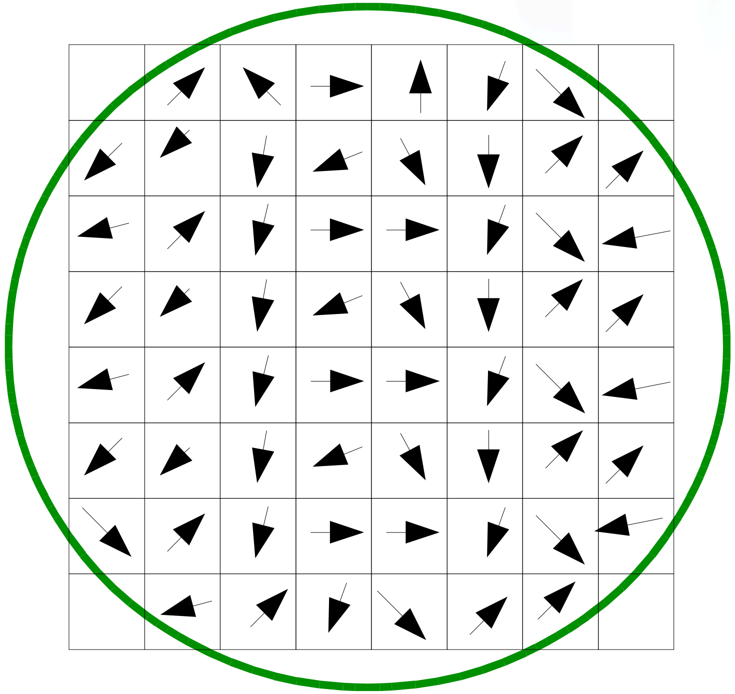
Must be as invariant as possible to changes in illumination and 3D viewpoint.

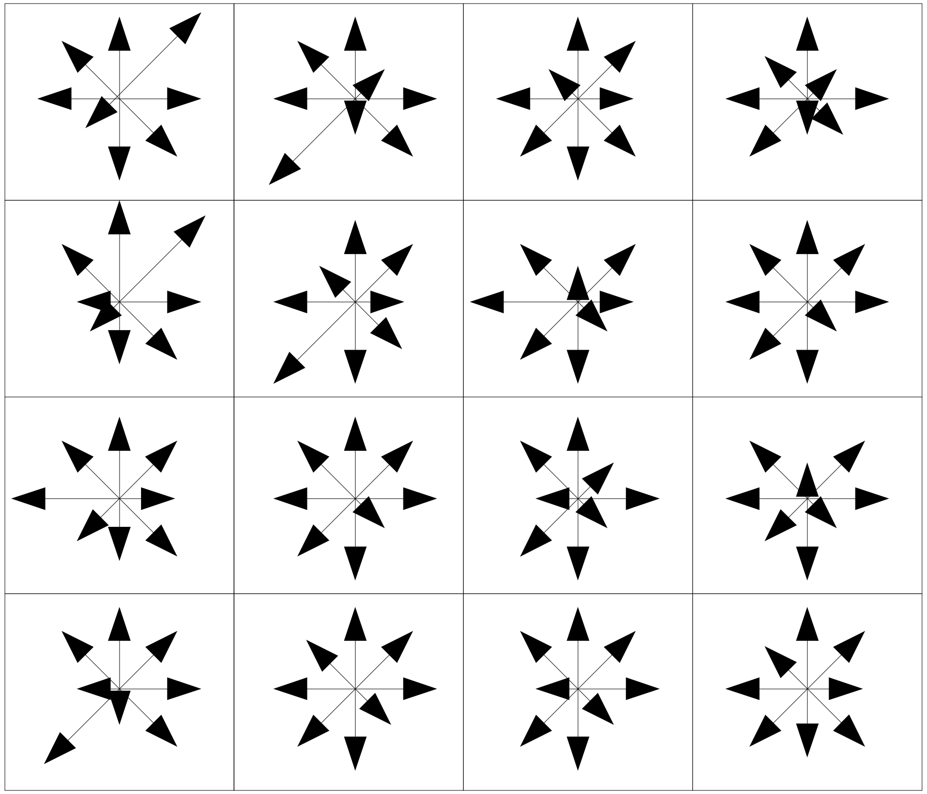
Set of orientation histograms are computed on 4x4 pixel areas.

Each gradient histogram contains 8 bins and each descriptor contains an array of 4 histograms.

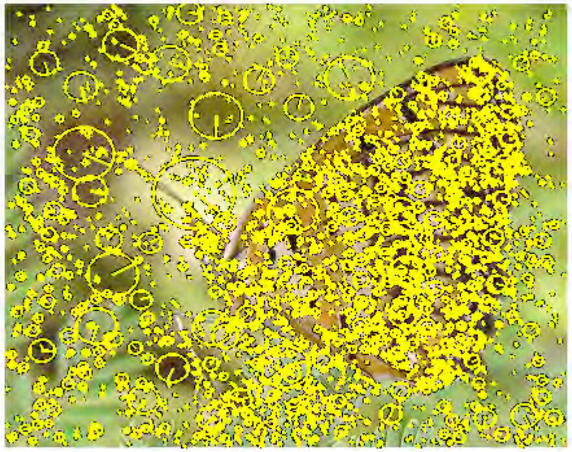
→ SIFT descriptor as 128 (4 x 4 x 8) element histogram

Visualising the keypoint descriptor





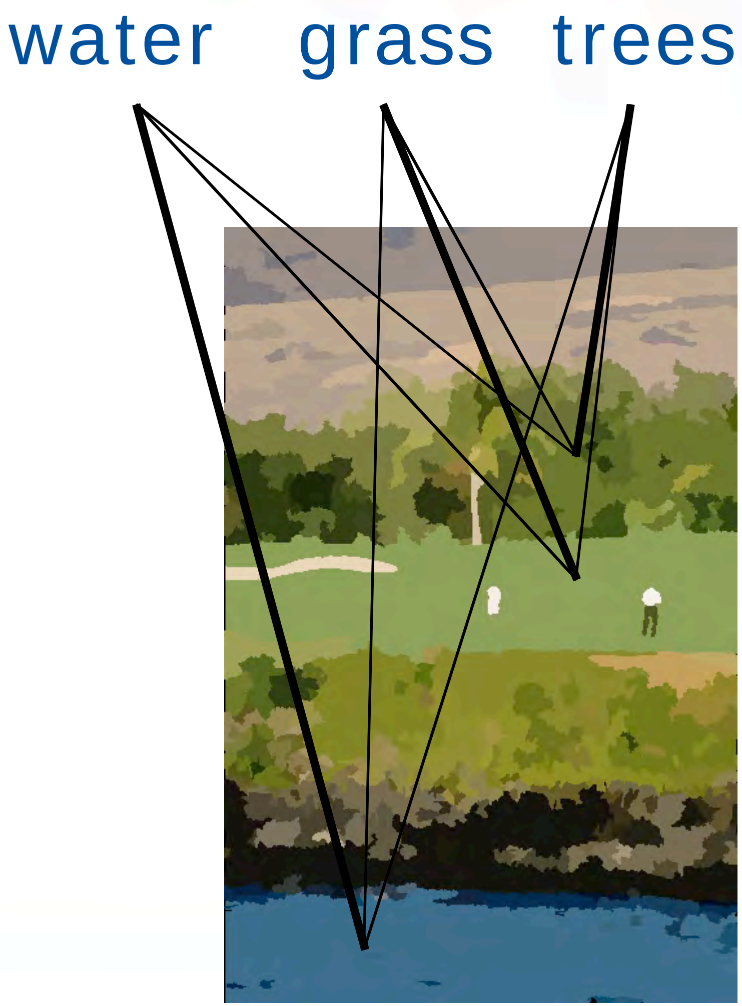
Example SIFT keypoints

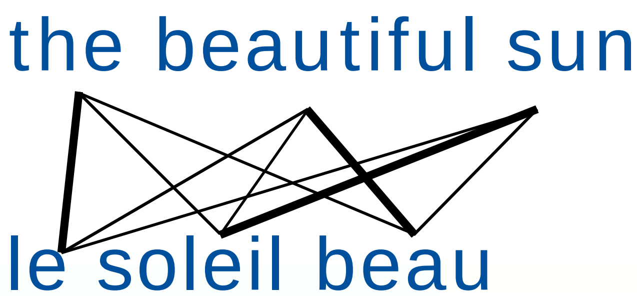


Multimedia information retrieval

1. What is multimedia information retrieval?
2. Metadata and piggyback retrieval
3. Multimedia fingerprinting
4. Automated annotation
5. Content-based retrieval

Automated annotation as machine translation





KNOWLEDGE MEDIA KMi INSTITUTE

Probabilistic models:

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

[with Magalhães, SIGIR 2005]
[with Yavlinsky and Schofield, CIVR 2005]
[with Yavlinsky, Heesch and Pickering: ICASSP May 2004]
[with Yavlinsky et al CIVR 2005]
[with Yavlinsky SPIE 2007]
[with Magalhães CIVR 2007, *best paper*]

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)

Automated annotation

[with Yavlinsky et al CIVR 2005]
[with Yavlinsky SPIE 2007]
[with Magalhaes CIVR 2007, best paper]

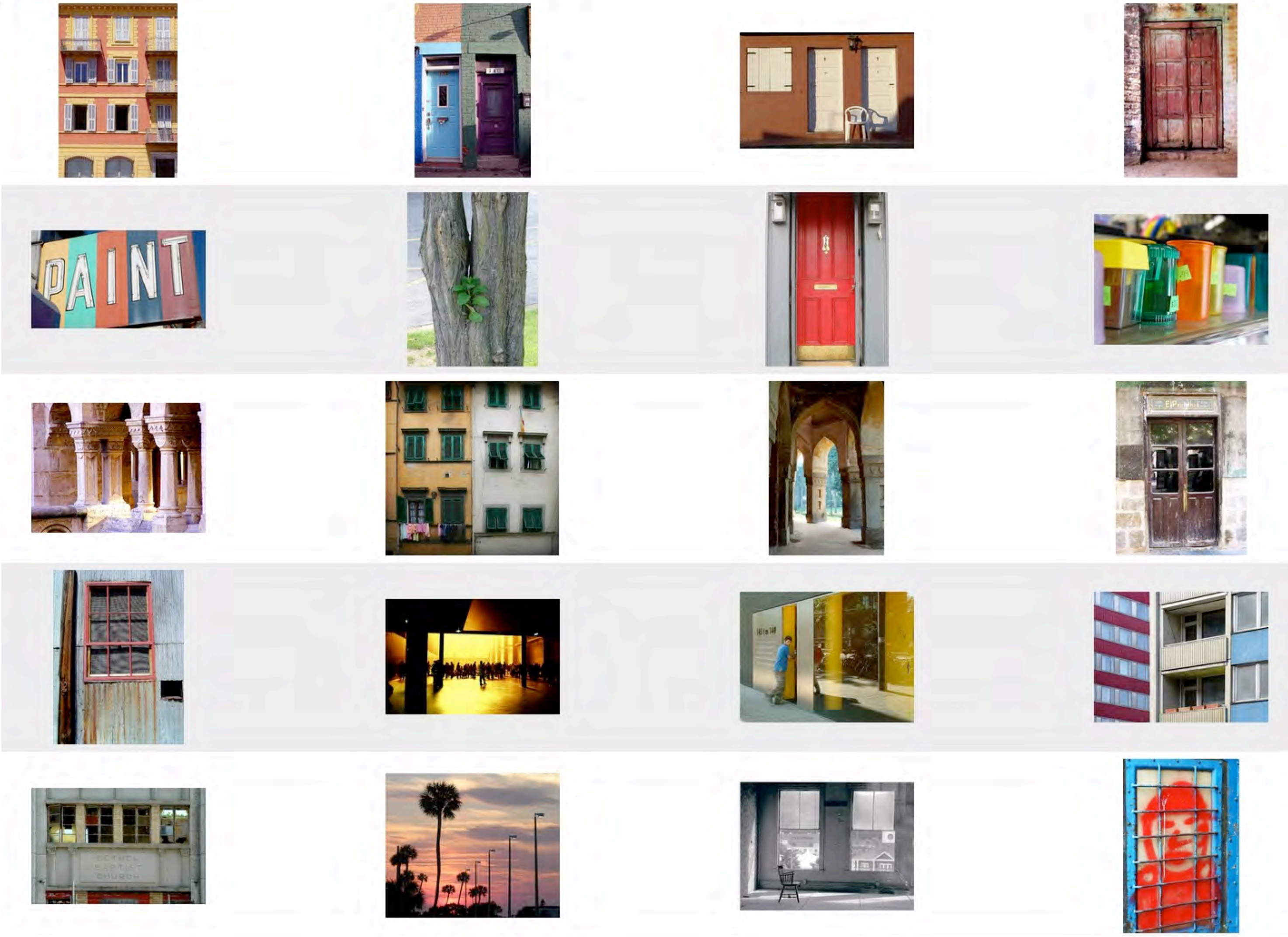
Automated: water buildings city sunset aerial

[Corel Gallery 380,000]

KNOWLEDGE MEDIA
KMi
INSTITUTE

The good

door



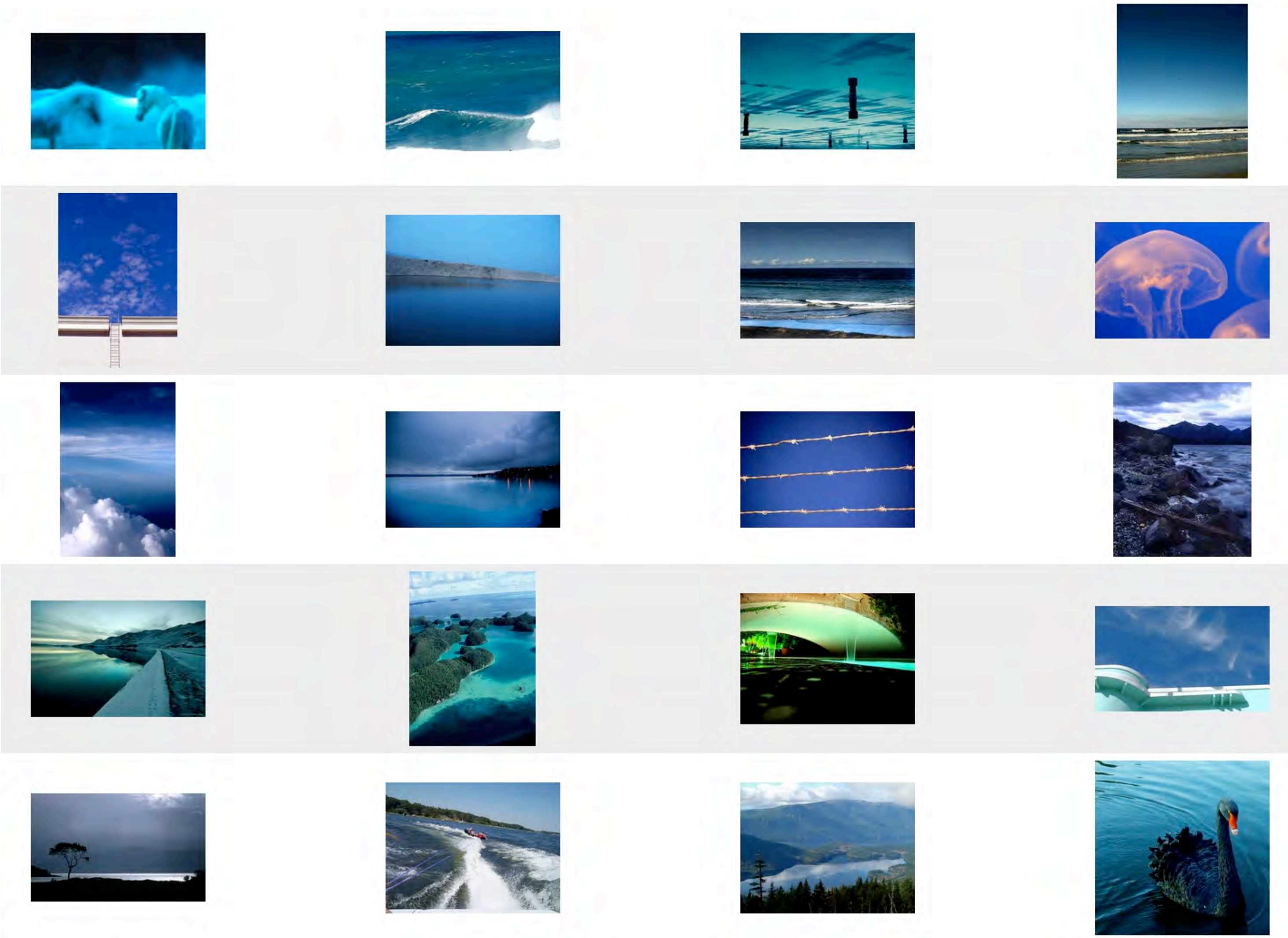
[beholdsearch.com, 19.07.2007,
now behold.cc (Yavlinksy)]
[images: Flickr creative commons]

The Open University

KNOWLEDGE MEDIA
KMi
INSTITUTE

The bad

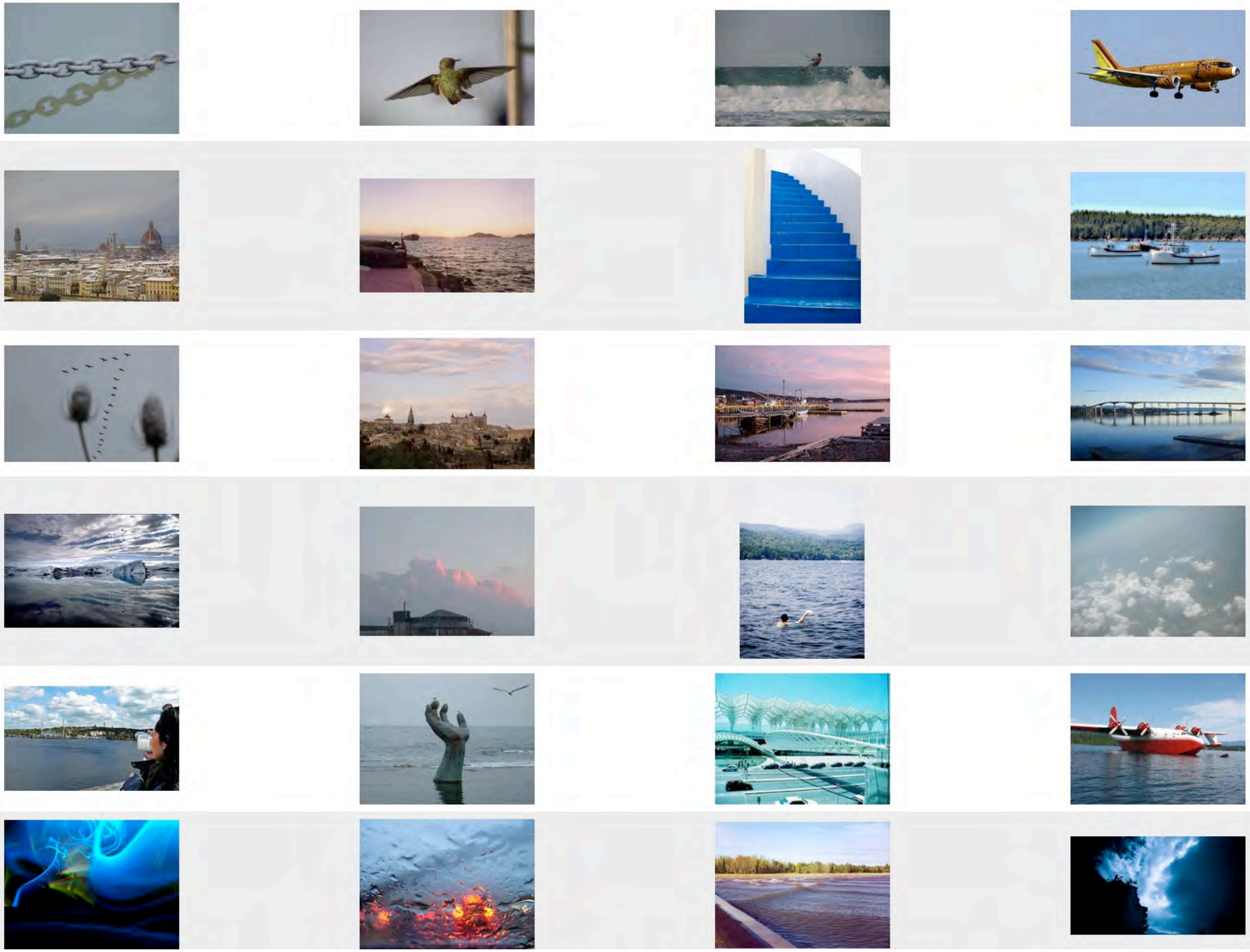
wave



[beholdsearch.com, 19.07.2007,
now behold.cc (Yavlinksy)]
[images: Flickr creative commons]

The ugly

iceberg



[beholdsearch.com, 19.07.2007,
now behold.cc (Yavlinksy)]
[images: Flickr creative commons]

Multimedia information retrieval

1. What is multimedia information retrieval?
2. Metadata and piggyback retrieval
3. Multimedia fingerprinting
4. Automated annotation
5. Content-based retrieval

Why content-based?

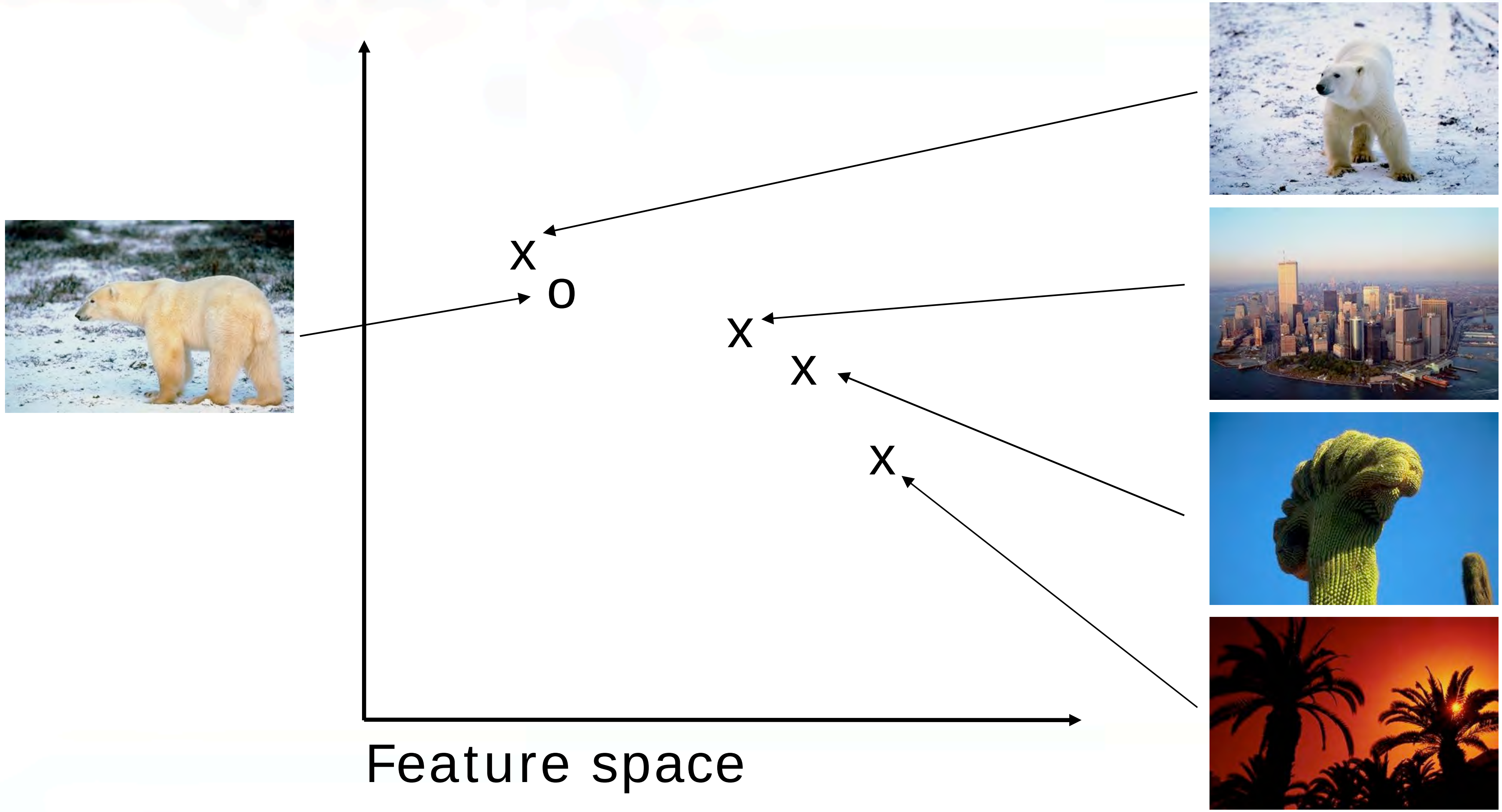
Give examples where we remember details by

- metadata?
- context?
- content (eg, “x” belongs to “y”)?

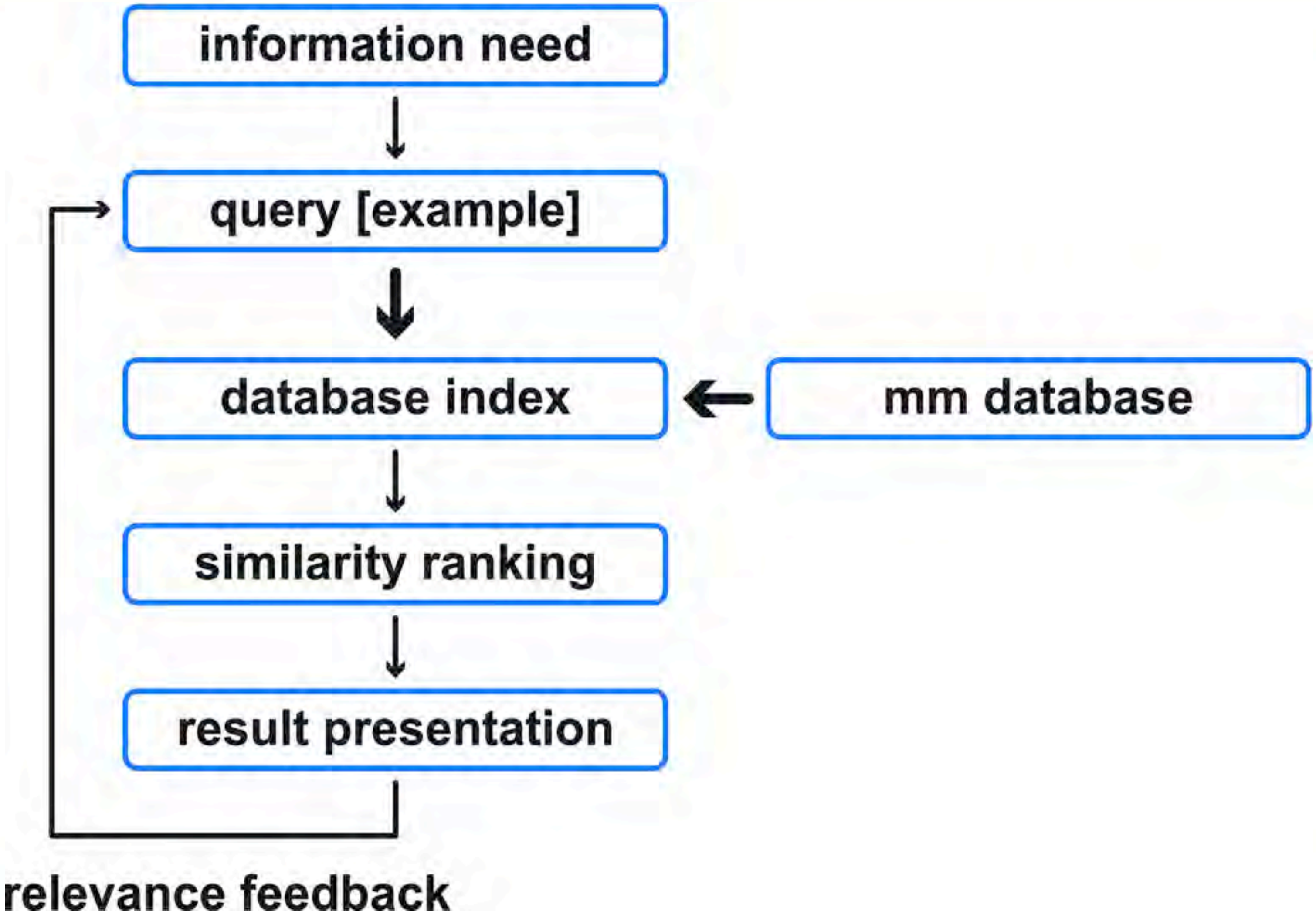
Metadata versus content-based: pro and con

-
-
-
-

Content-based retrieval: features and distances



Content-based retrieval: Architecture



Features

Visual

- Colour, texture, shape, edge detection, SIFT/SURF

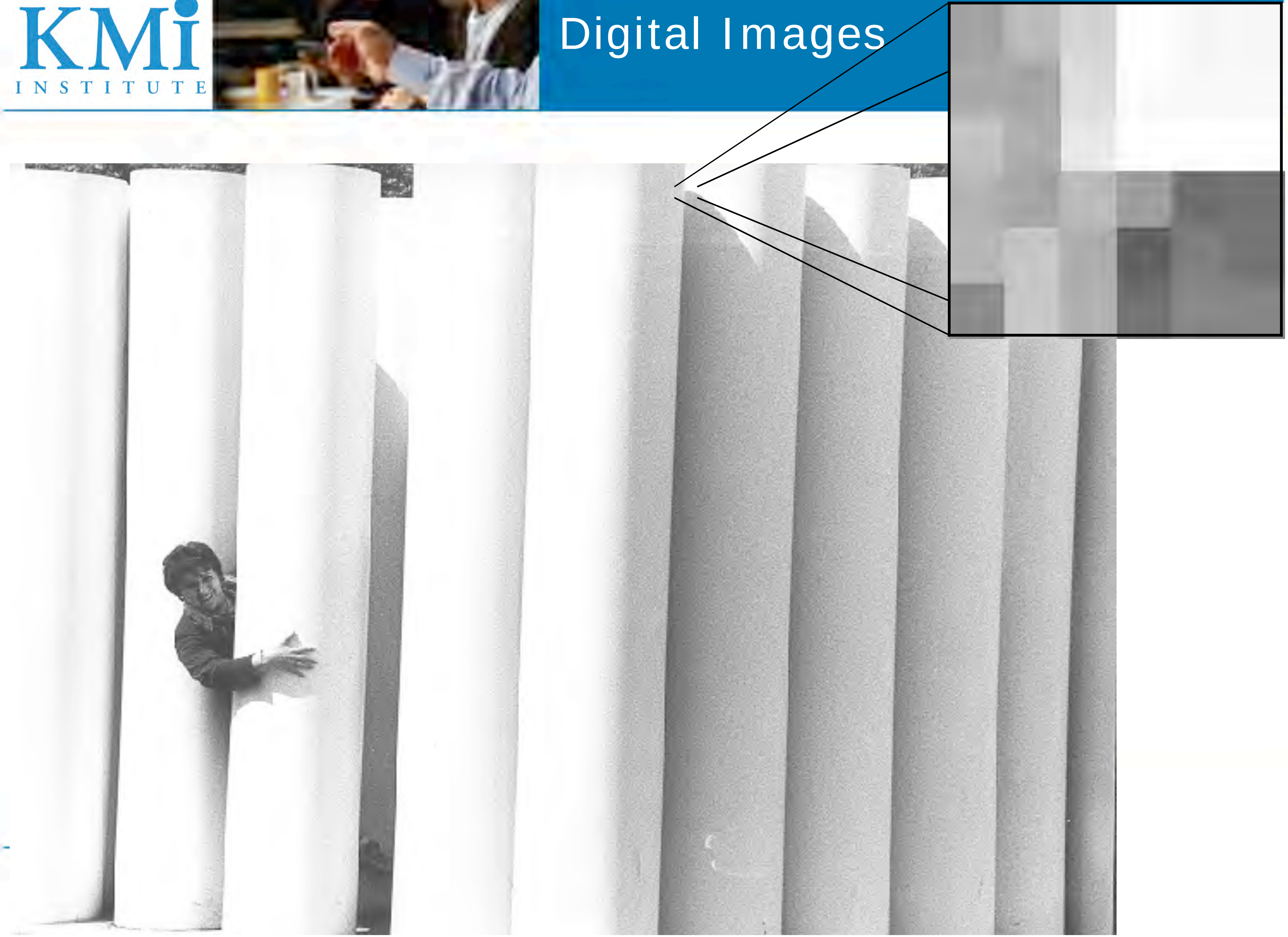
Audio

Temporal

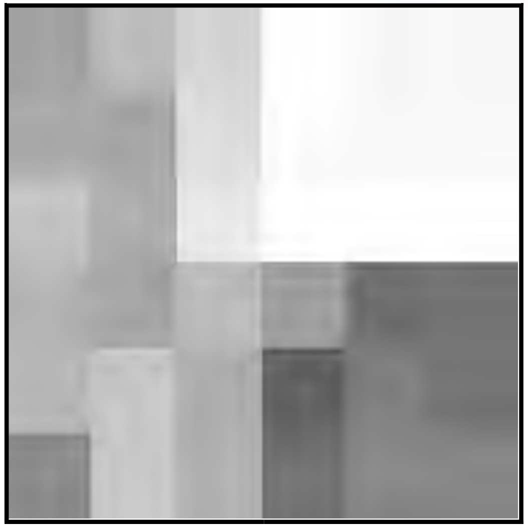
How to describe the features?

- For people
- For computers

Digital Images

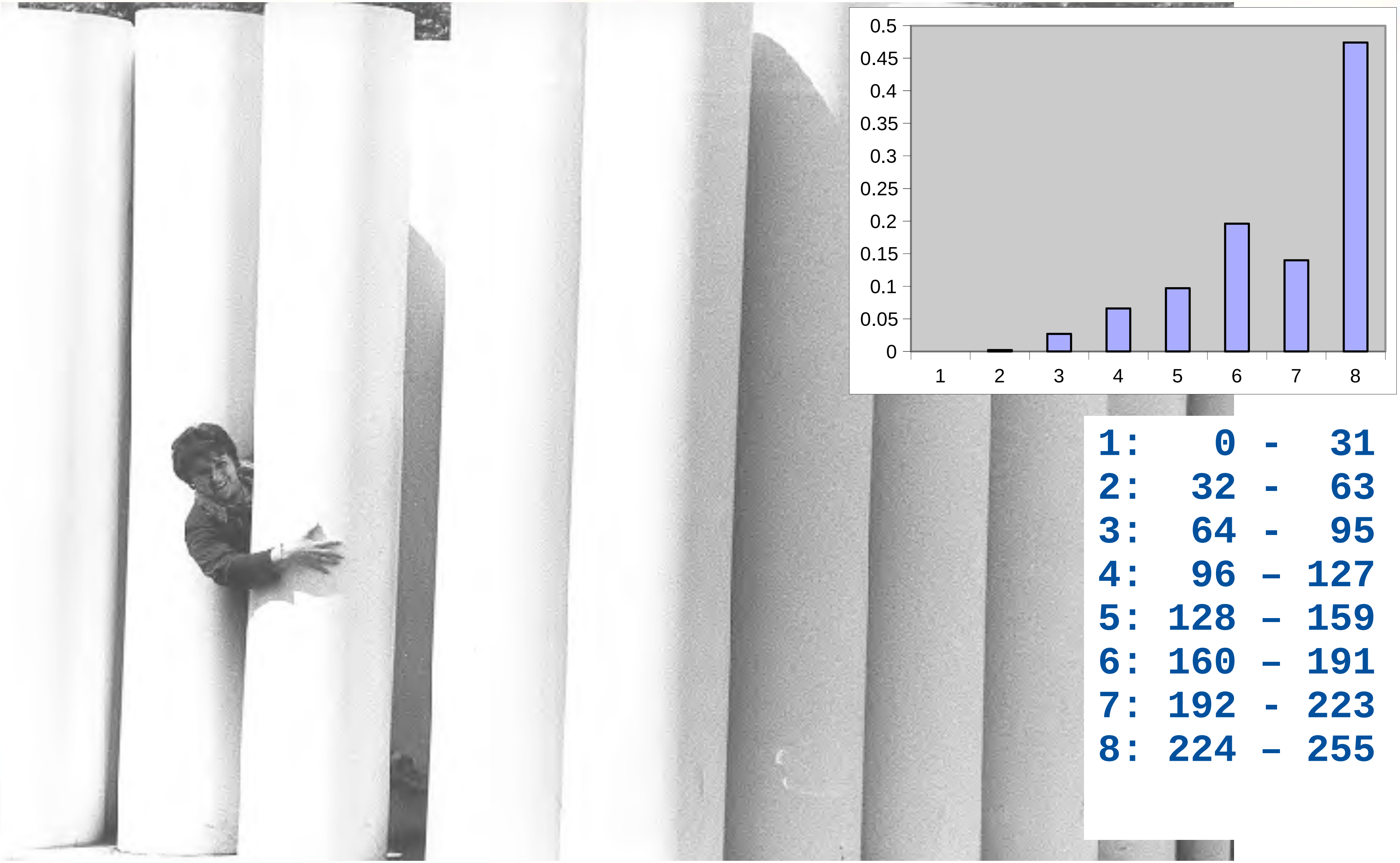


Content of an image



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



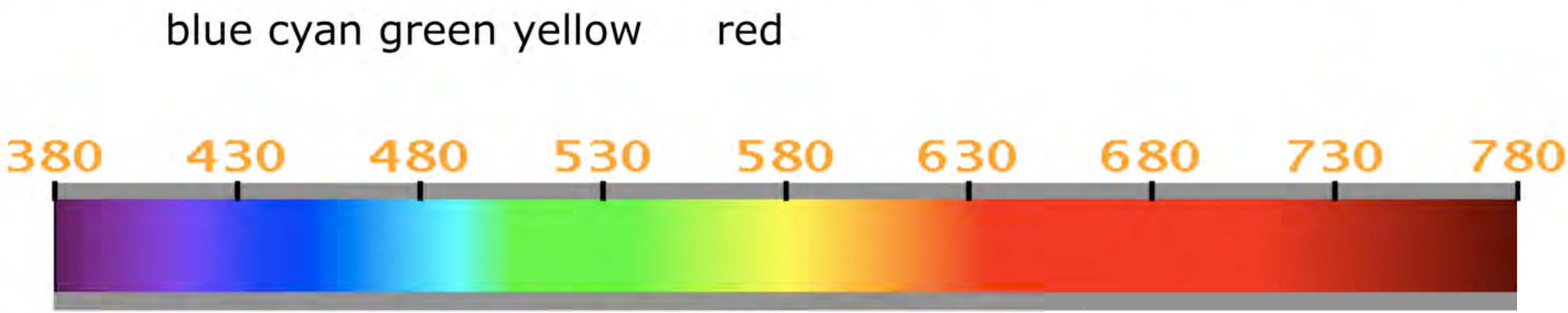
1: 0 - 31
2: 32 - 63
3: 64 - 95
4: 96 – 127
5: 128 – 159
6: 160 – 191
7: 192 - 223
8: 224 – 255

Colour

phenomenon of human perception

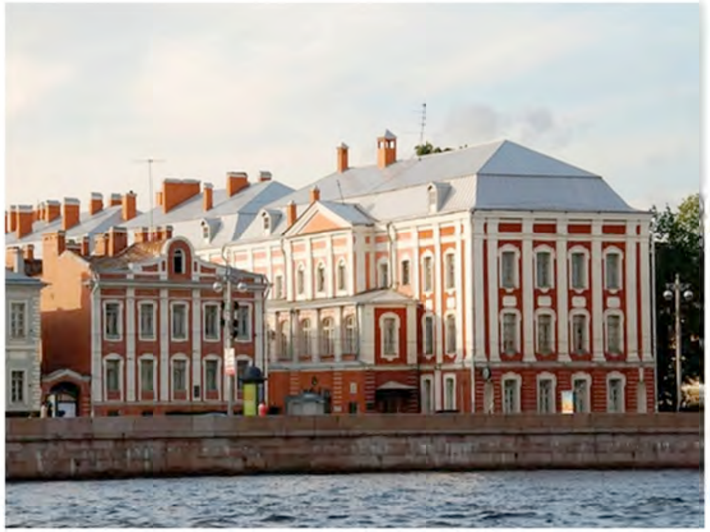
three-dimensional (RGB/CMY/HSB)

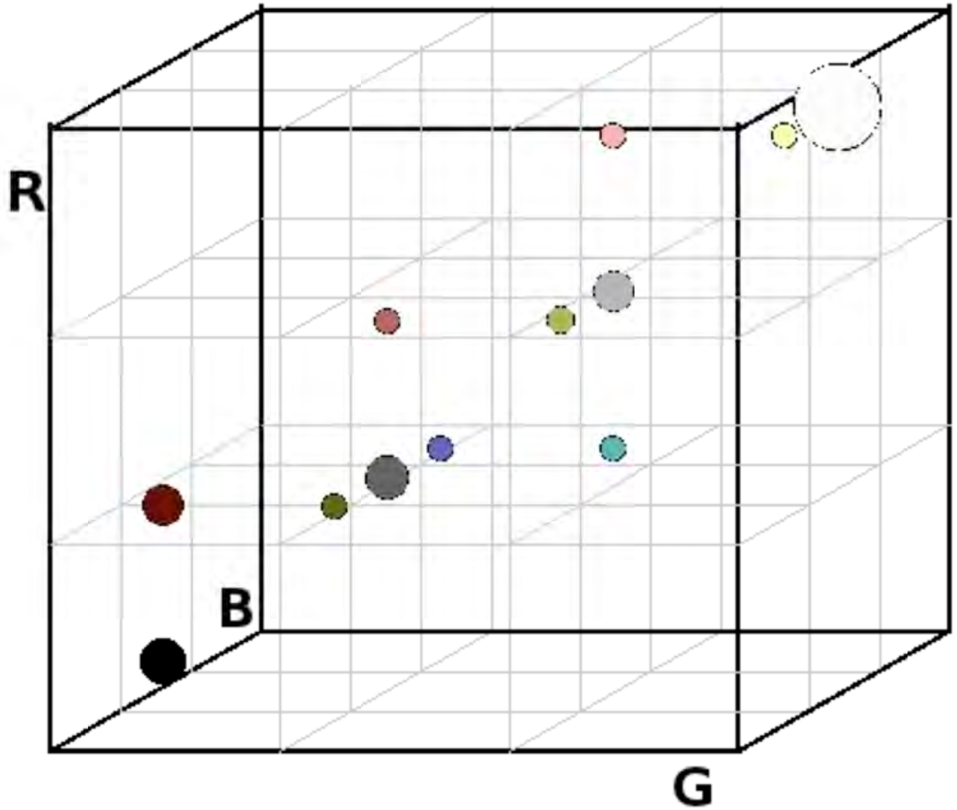
spectral colour: pure light of one wavelength



spectral colours: wavelength (nm)

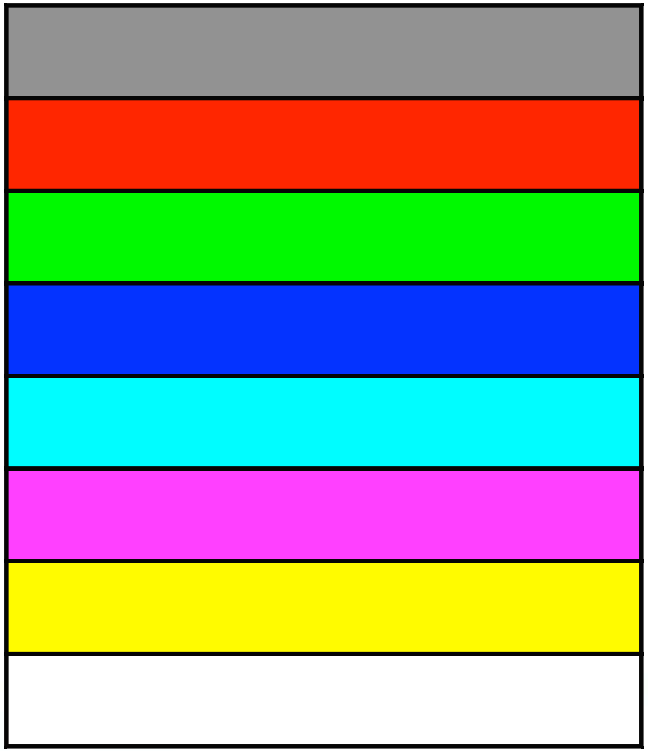
Colour histogram



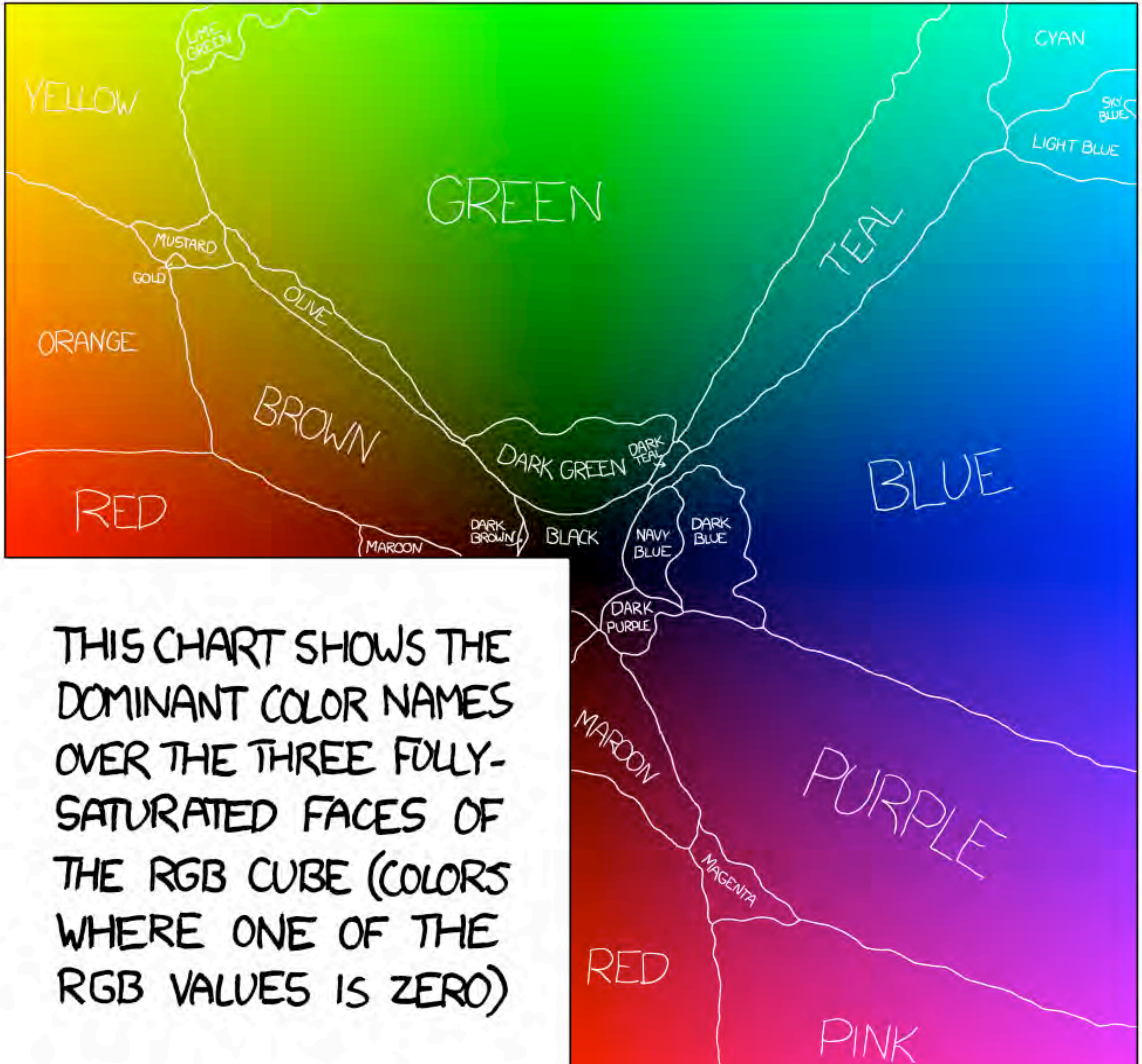


Exercise

Sketch a 3D colour histogram for



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



http://blog.xkcd.com/2010/05/03/color-survey-

HSB colour model

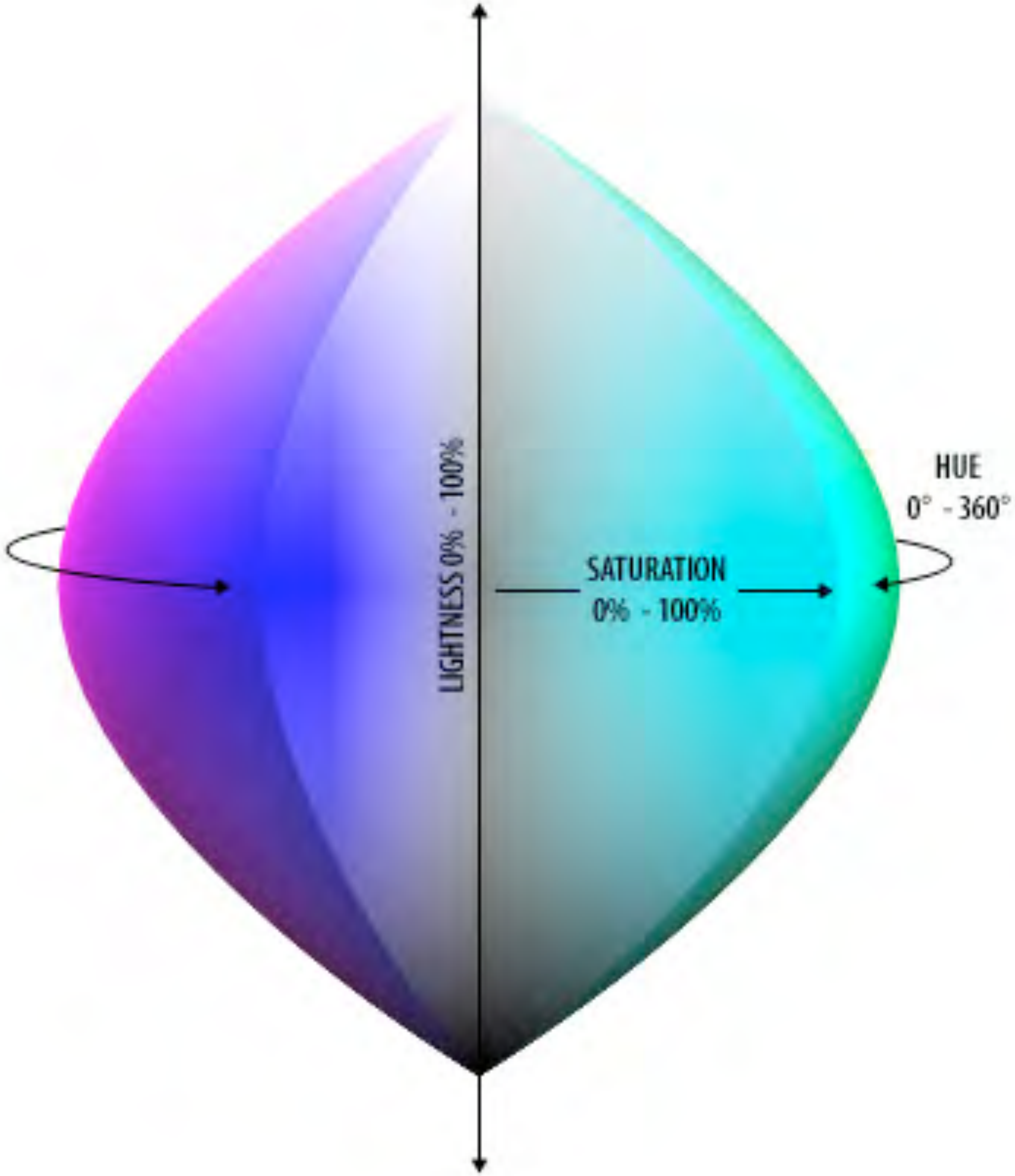
saturation (0% - 100%)
= spectral purity

hue (0°-360°)
spectral colour

brightness (0% - 100%)
= energy or luminance

chromaticity = hue+saturation

HSB colour model



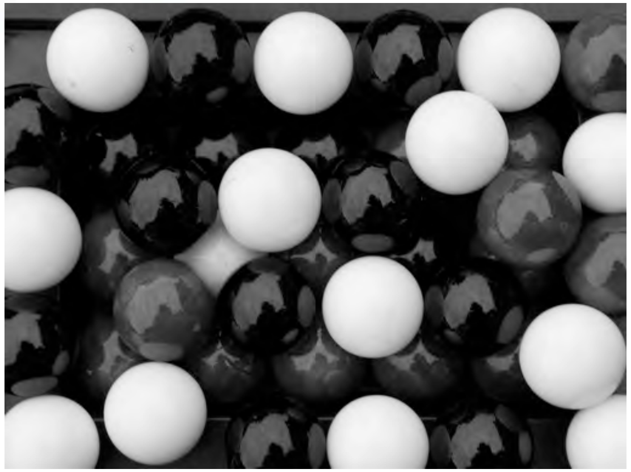
HSB model

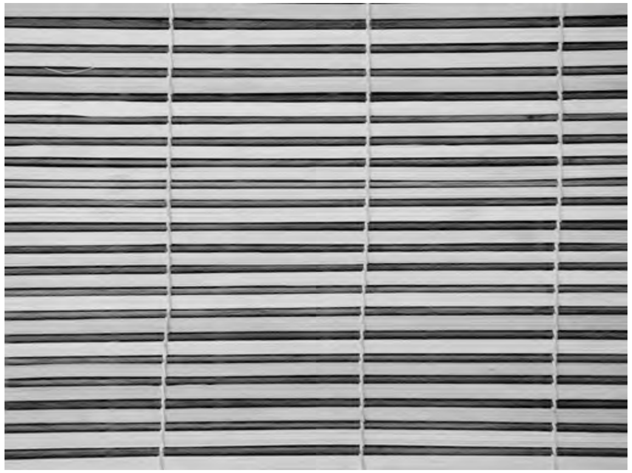
disadvantage: hue coordinate is not continuous

- 0 and 360 degrees have the same meaning
- but there is a huge difference in terms of numeric distance
- example:
  red = (0°,100%,50%) = (360°,100%,50%)

advantage: it is more natural to describe colour changes “brighter blue”, “purer magenta”, etc

Texture



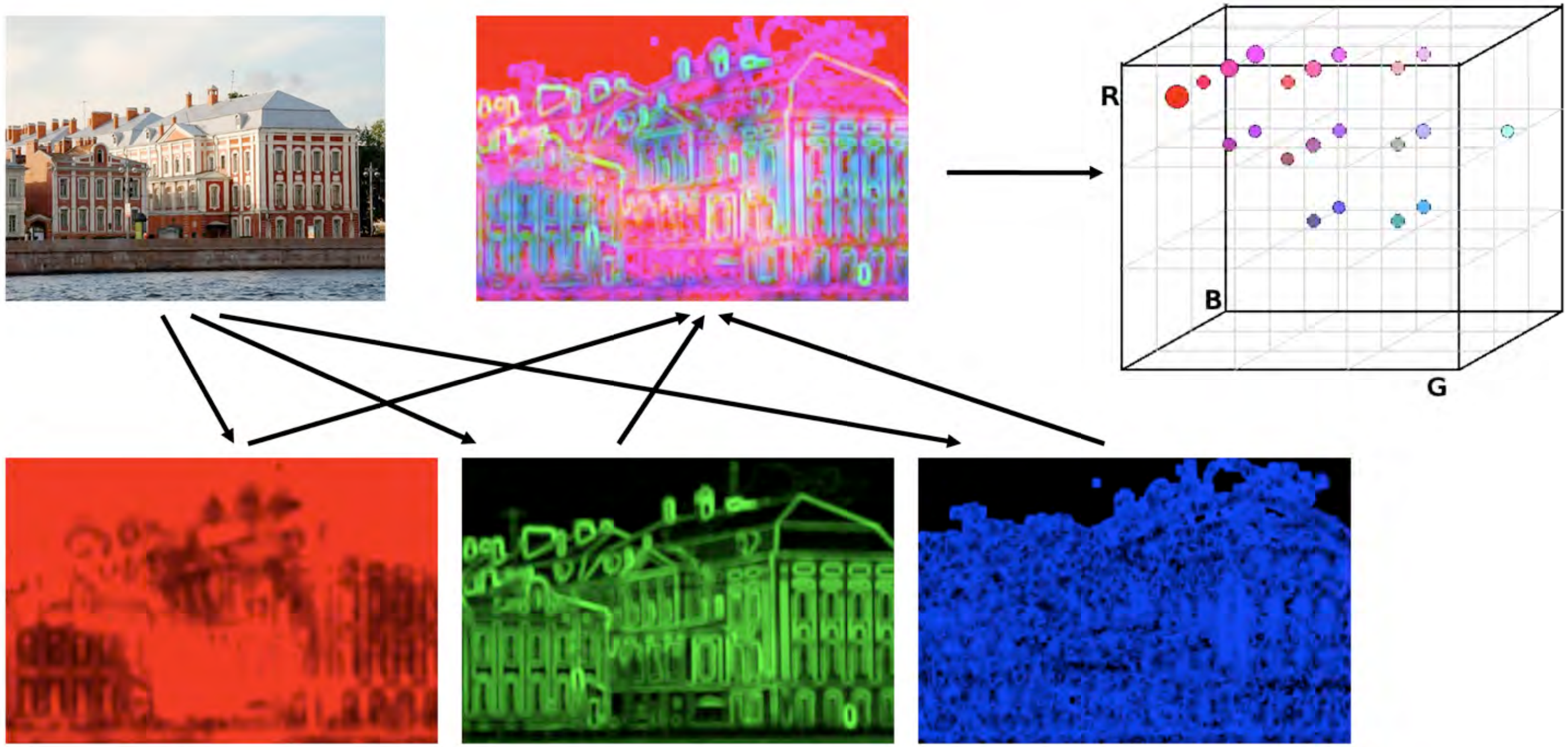


coarseness

contrast

directionality

Texture histograms

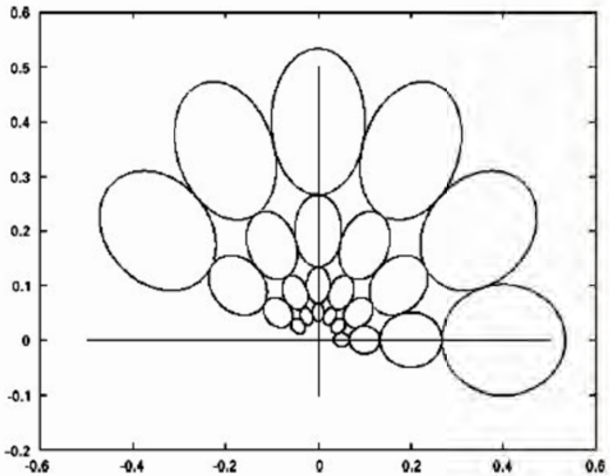


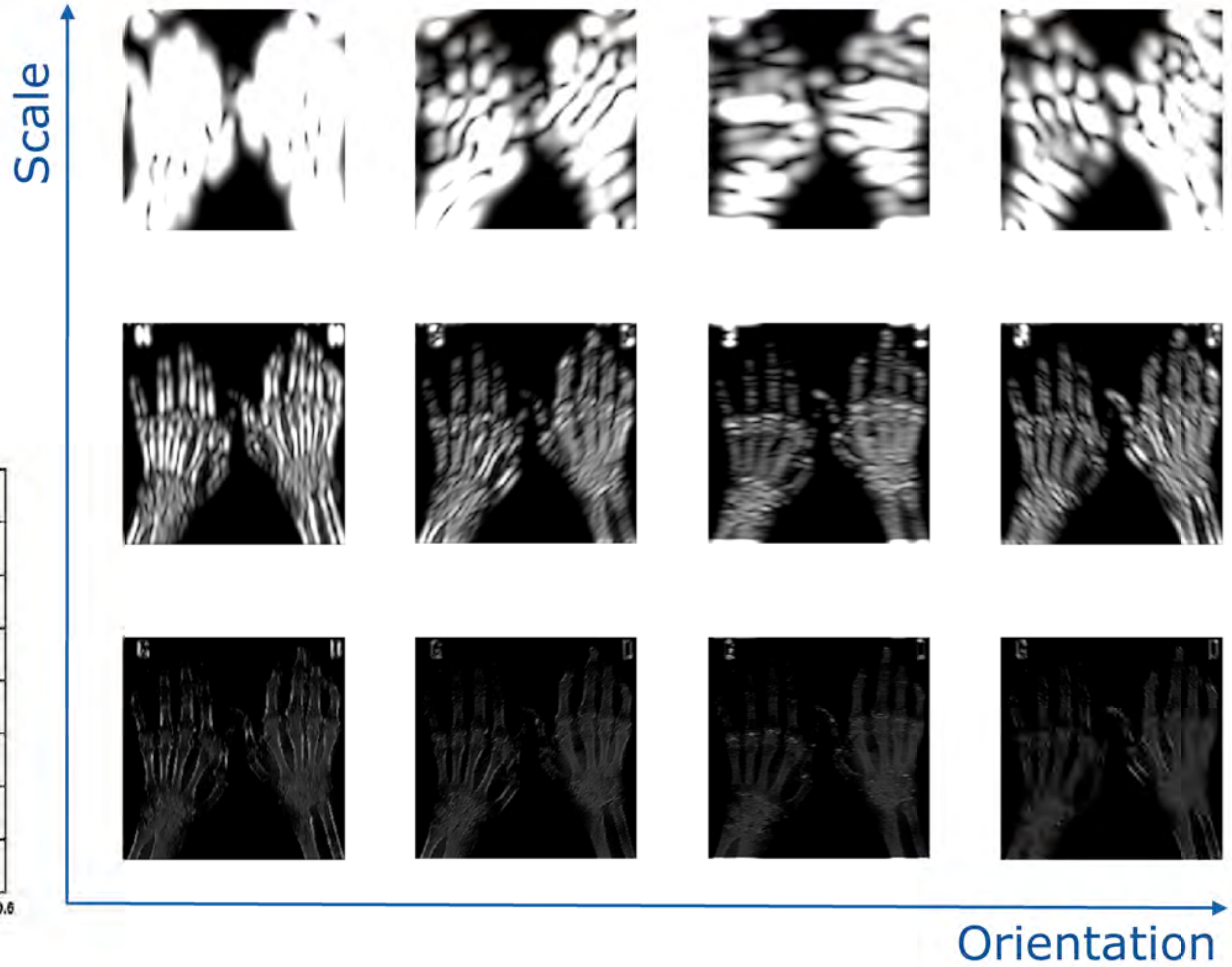
Coarseness coNtrast Directionality

[with Howarth, *IEE Vision, Image & Signal Proc* 15(6) 2004; Howarth PhD thesis]

Gabor filter

Query





[with Howarth, CLEF 2004]

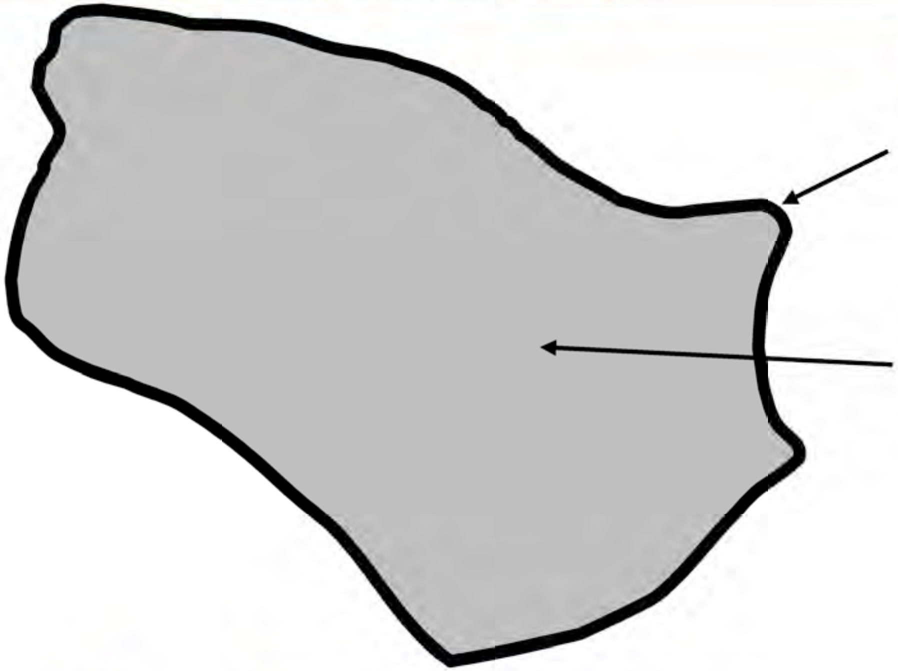
Shape Analysis

shape = class of geometric objects invariant under
- translation
- scale (changes keeping the aspect ratio)
- rotations

information preserving description
(for compression)

non-information preserving (for retrieval)
- boundary based (ignore interior)
- region based (boundary+interior)

Perimeter and area



$$P = \int \sqrt{x'^2(t) + y'^2(t)}\,dt$$

~~$$A = \iint_R dxdy$$~~

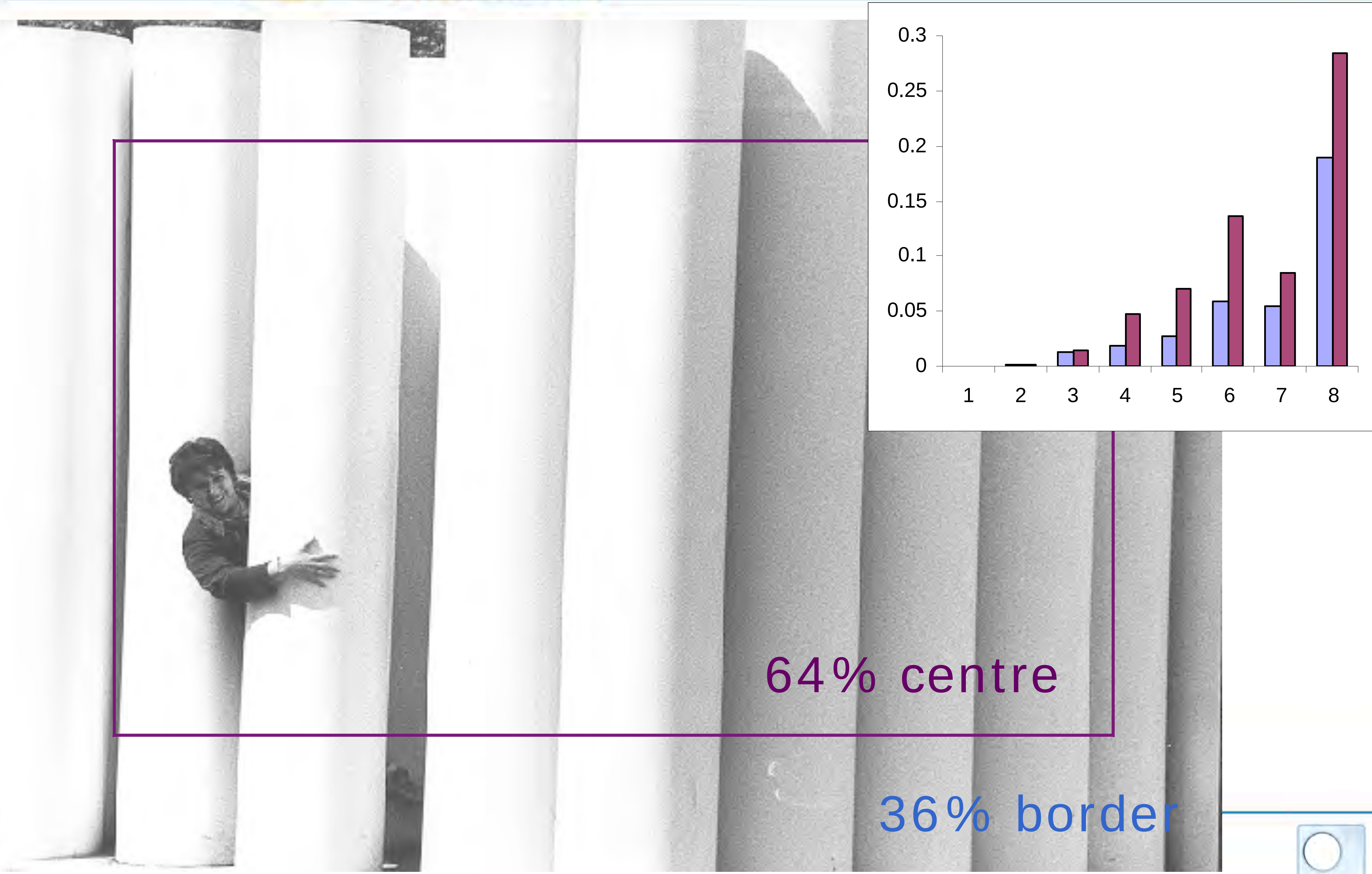
~~boundary pixel count~~

count pixels in area

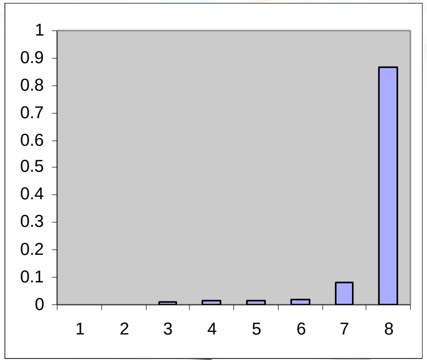
Global vs local

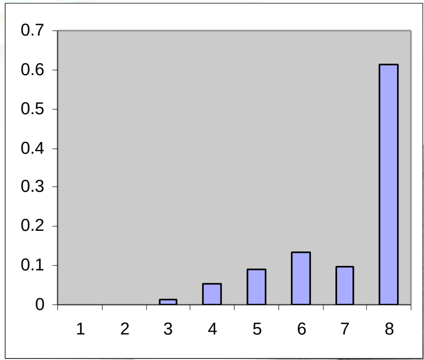
Global histogram also matches polar bears, marble floors, …

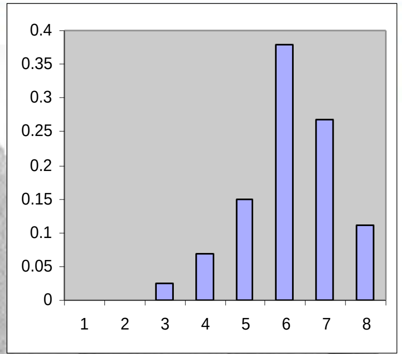
Localisation

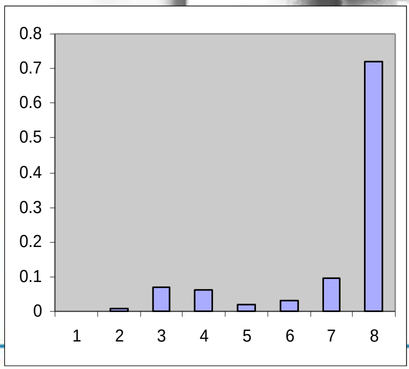


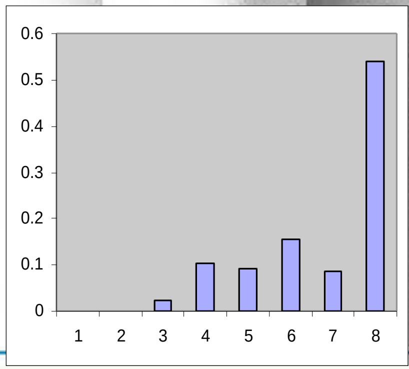
Tiled Histograms

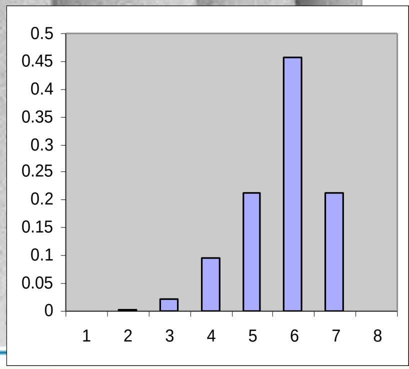




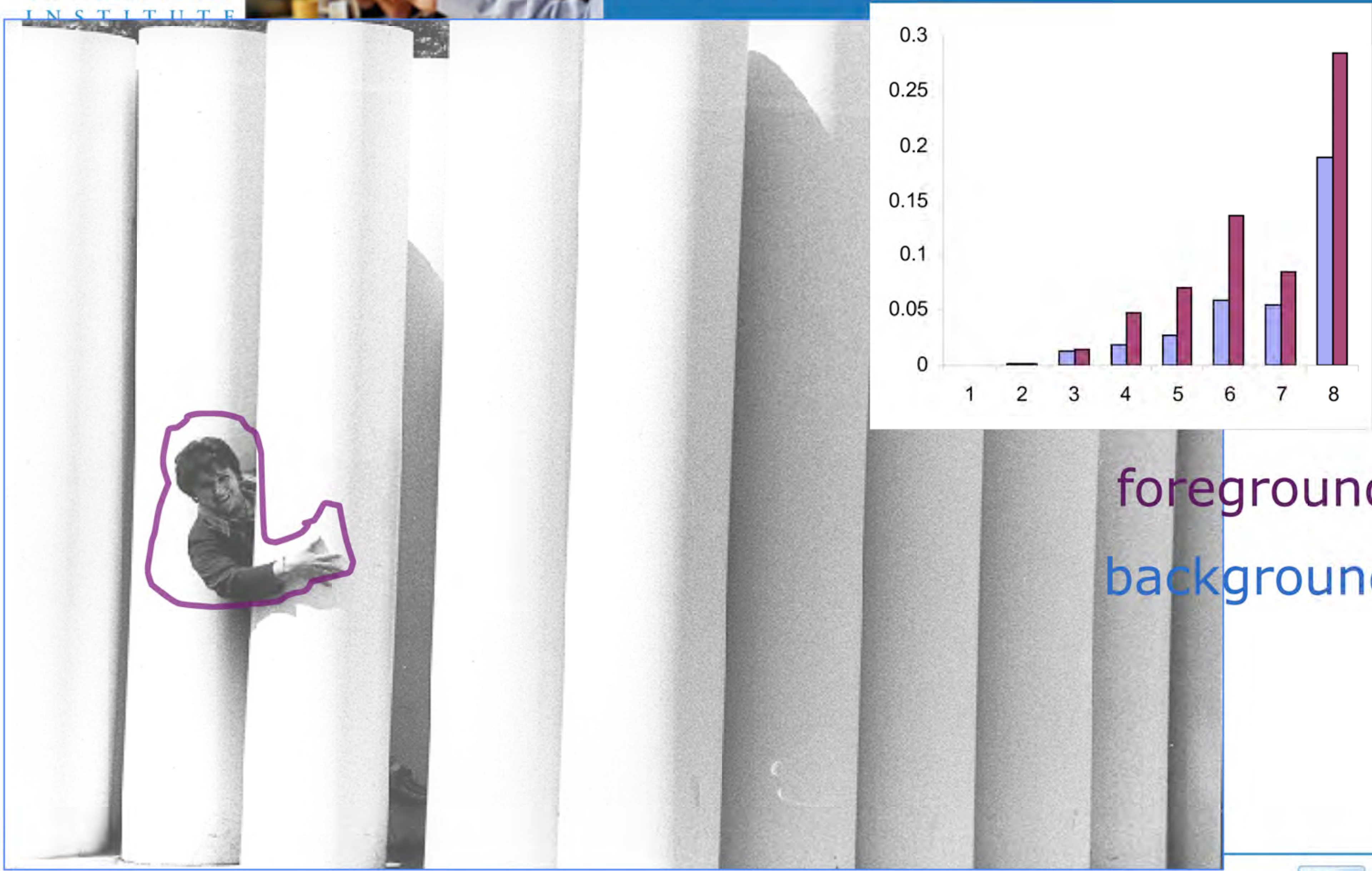








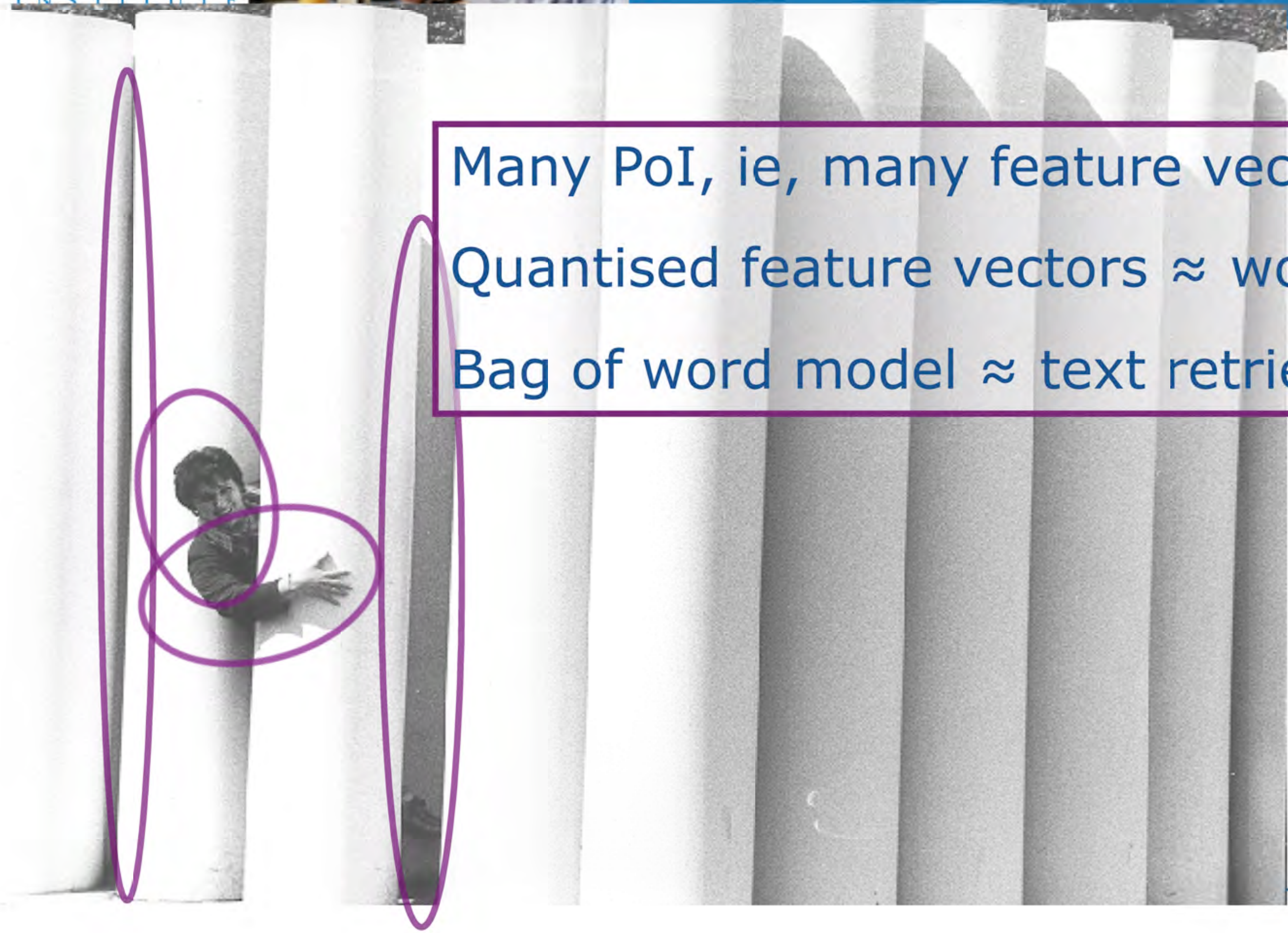
Segmentation



foreground

background

Points of interest



Many PoI, ie, many feature vectors

Quantised feature vectors ≈ words

Bag of word model ≈ text retrieval

Video Segmentation

gradual transition detection (eg, fade)
- accumulate distances
- long-range comparison

audio cues
- silence and/or speaker change

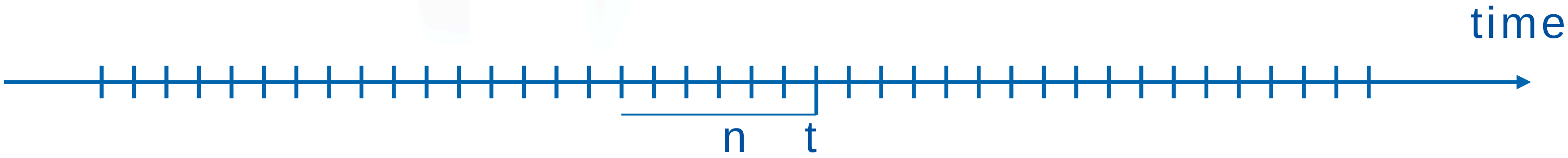
motion detection and analysis
- camera motion, zoom, object motion
- MPEG provides some motion vectors

[Vlad Tanasescu: Anticipation, SCiFi trailer]]

[Vlad Tanasescu: Anticipation, SCiFi trailer]]

Long range comparison

At time t define distance $d_n(t)$



- compare frames t-n+i and t+i (i=0,...,n-1)

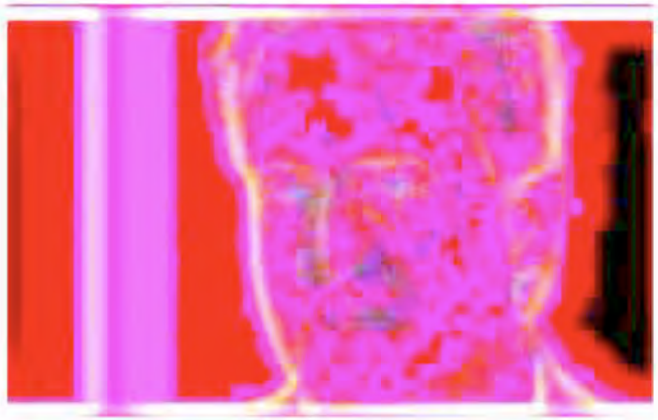
- average their respective distances over i

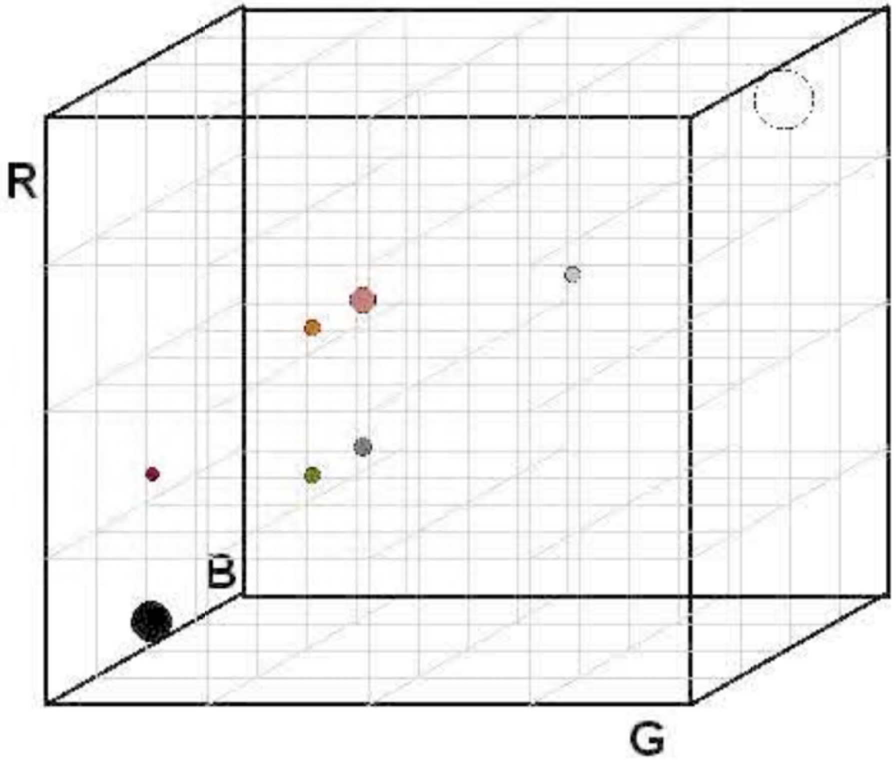
Peak in $d_n(t)$ detected if

$d_n(t) > \text{threshold}$ and

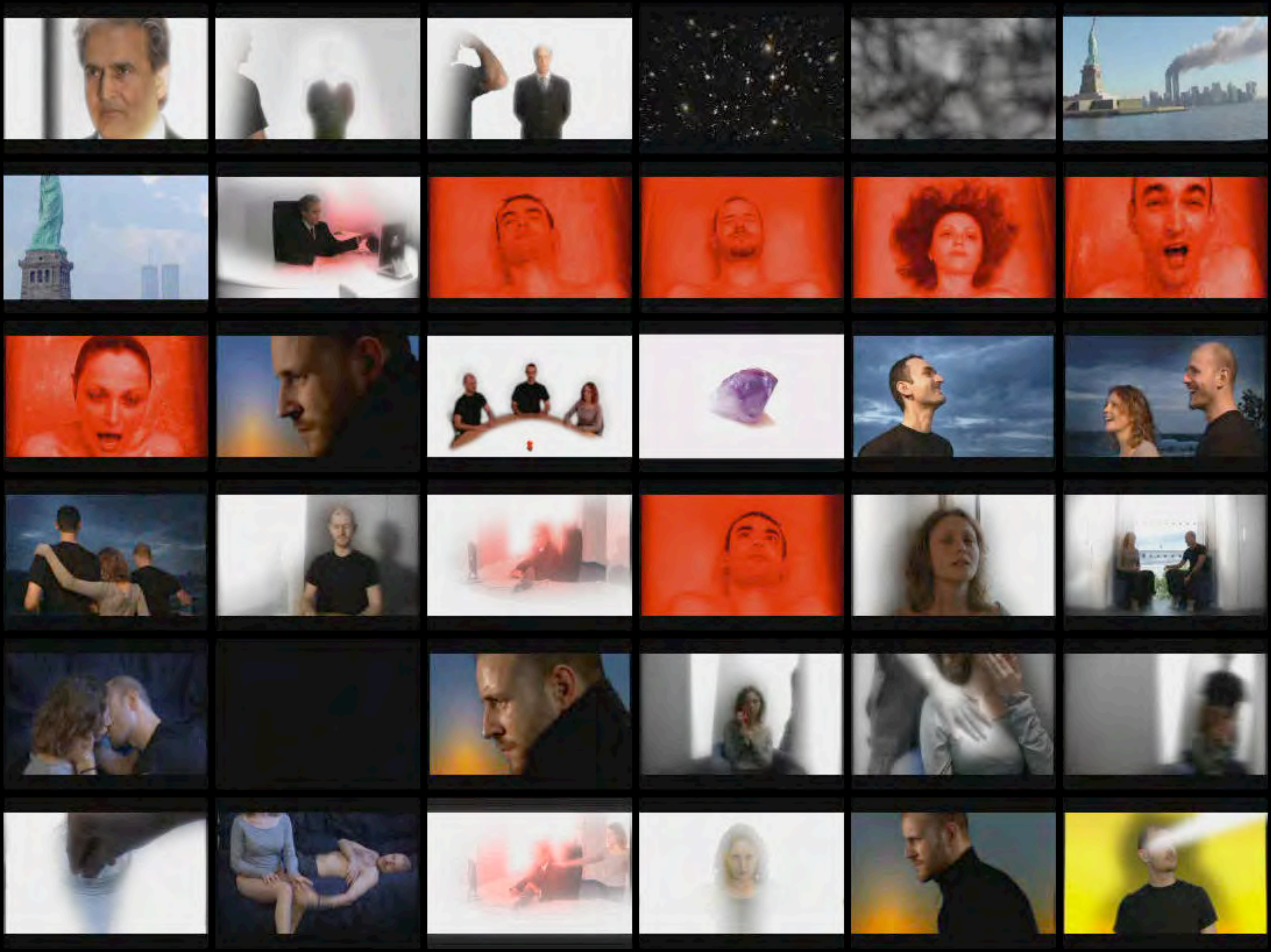
$d_n(t) > d_n(s)$ for all neighbouring s

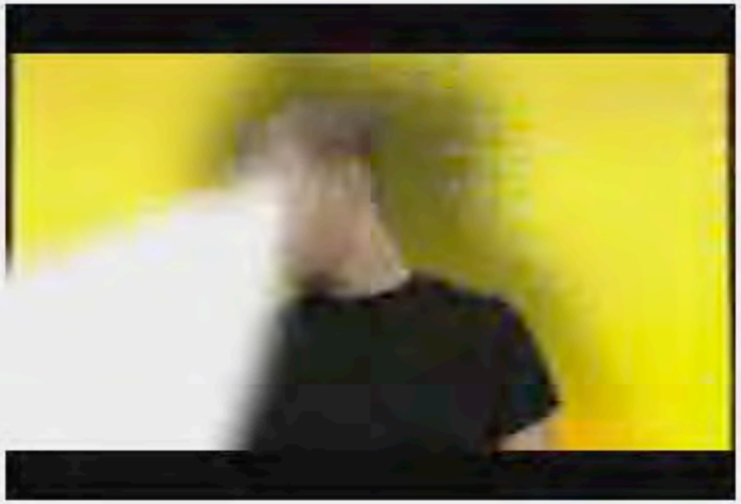
Shot = near-coincident peaks of d_{16} and d_8

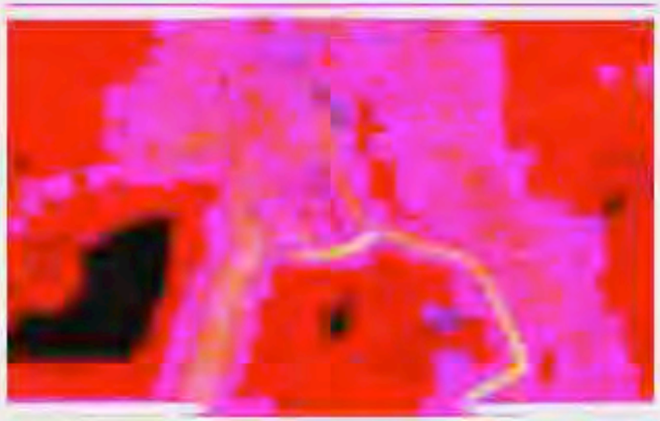


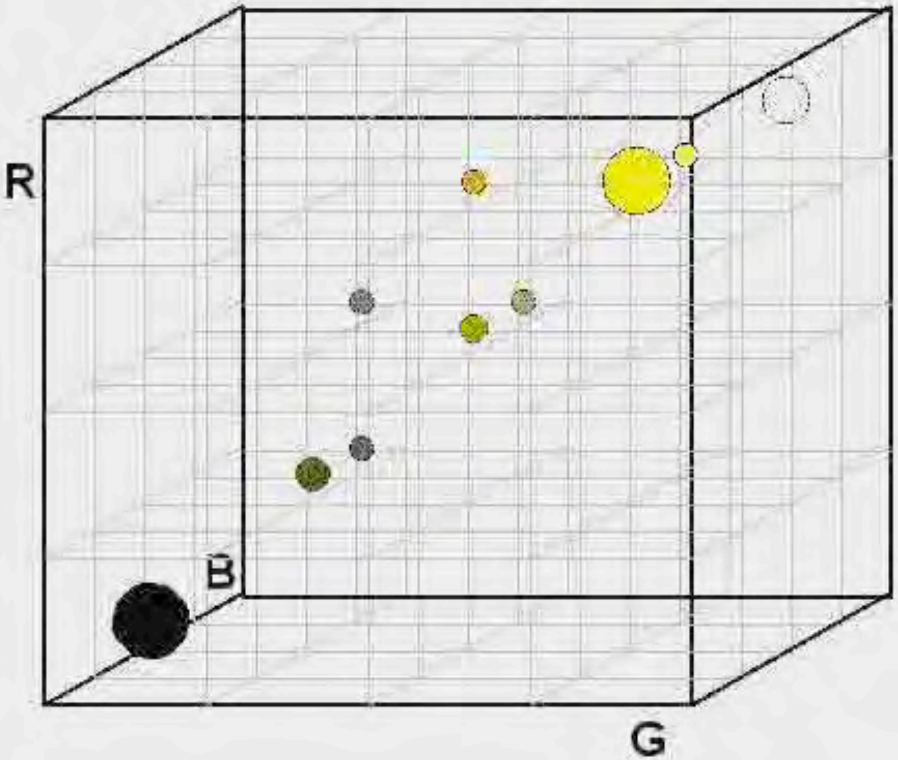

R
B
G

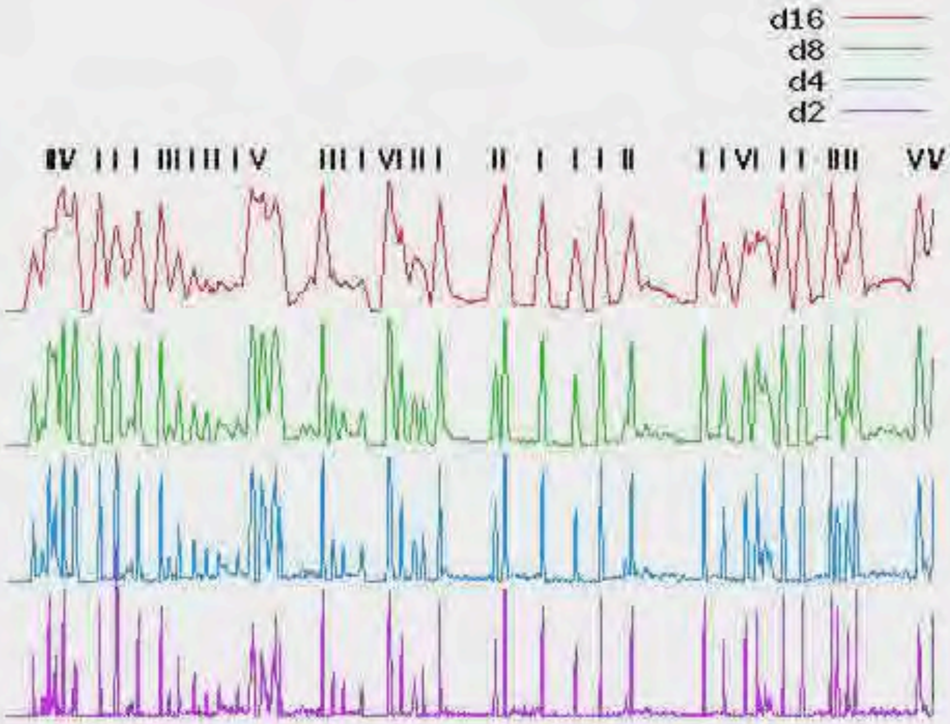

d16
d8
d4
d2





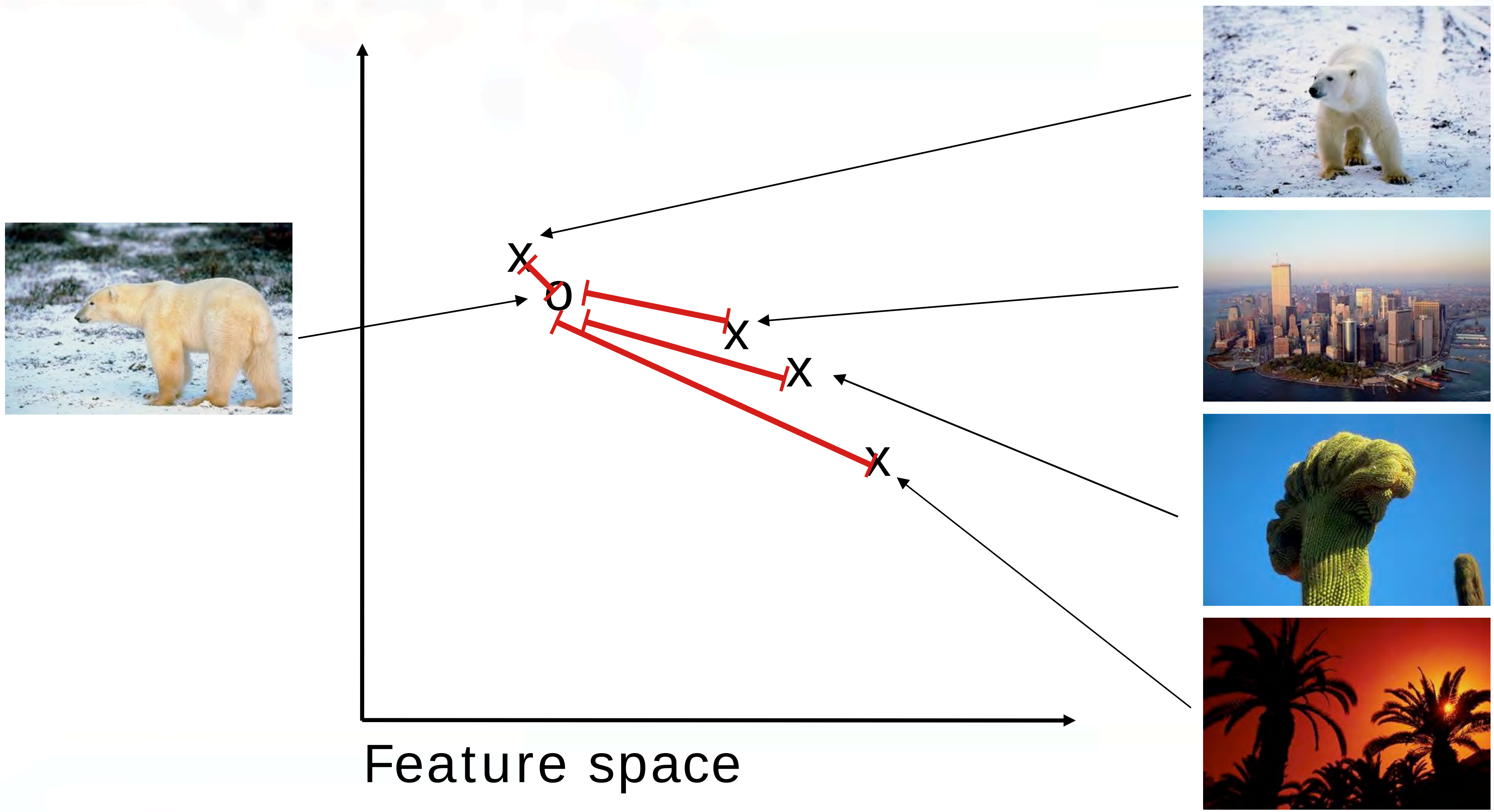



R
B
G


d16
d8
d4
d2

KNOWLEDGE MEDIA KMi INSTITUTE

Features and distances



Distances and similarities

- assumes coding of MM objects as data vectors
  - distance measures
    - Euclidean, Manhattan
  - correlation measures
    - Cosine similarity measure
    - histogram intersection for normalised histograms

$$\mathrm{sim}(h, q) = \sum_i \min(h_i, q_i)$$

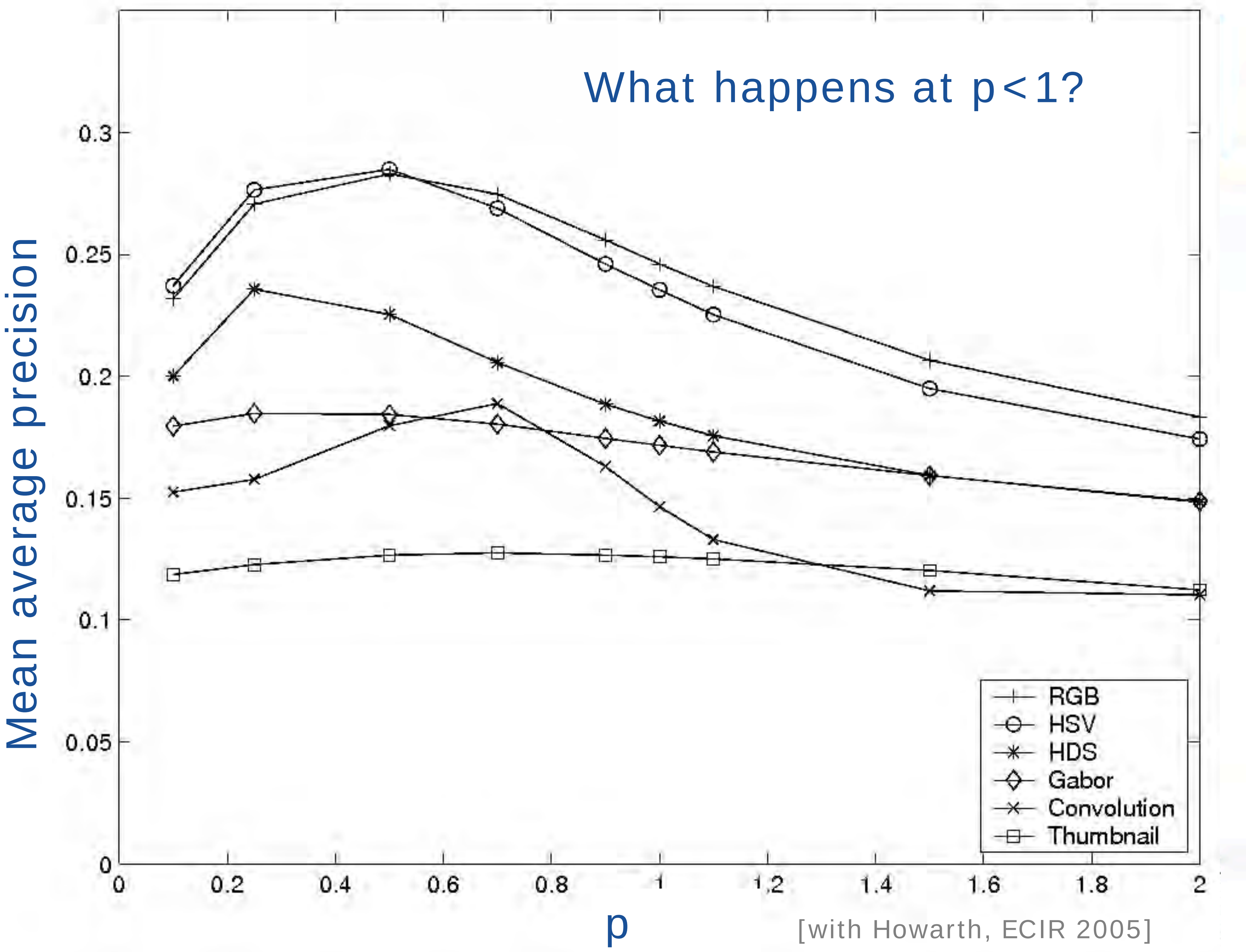
L_2 vs L_1

$$d_p(v, w) = \sqrt[p]{\sum_i |v_i - w_i|^p}, \quad p \geq 1$$

$$d_2(v, w) = \sqrt[2]{\sum_i |v_i - w_i|^2}$$

$$d_1(v, w) = \sqrt[1]{\sum_i |v_i - w_i|^1} = \sum_i |v_i - w_i|$$

$$d_\infty(v, w) =$$



[with Howarth, ECIR 2005]

Other distance measures

- Squared chord
- Earth Mover's Distance
- Chi squared distance
- Kullback-Leibler divergence (not a true distance)
- Ordinal distances (for string values)

Best distance?

Squared chord

$$d_{sc}(v, w) = \sum_{i=1}^{n} \left(\sqrt{v_i} - \sqrt{w_i}\right)^2$$

[with Liu et al, AIRS 2008; with Hu et al, ICME 2008]

Recap: Multimedia information retrieval

1. What is multimedia information retrieval?
2. Metadata and piggyback retrieval
3. Multimedia fingerprinting
4. Automated annotation
5. Content-based retrieval

KNOWLEDGE MEDIA
KMi
INSTITUTE

Multimedia Information Retrieval

Prof Stefan Rüger

Multimedia and Information Systems
Knowledge Media Institute
The Open University

http://kmi.open.ac.uk/mmis

MMIS
Multimedia and Information Systems