KNOWLEDGE MEDIA KMi INSTITUTE

Multimedia Information Retrieval

4th Russian Summer School in Information Retrieval, September 2010

RuSSIR
Russian Summer School in 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

Multimedia Information Retrieval

1 What is multimedia information retrieval?
2 Basic multimedia search technologies
3 Evaluation of MIR Systems
4 Added value

KNOWLEDGE MEDIA KMi INSTITUTE

Multimedia Information Retrieval

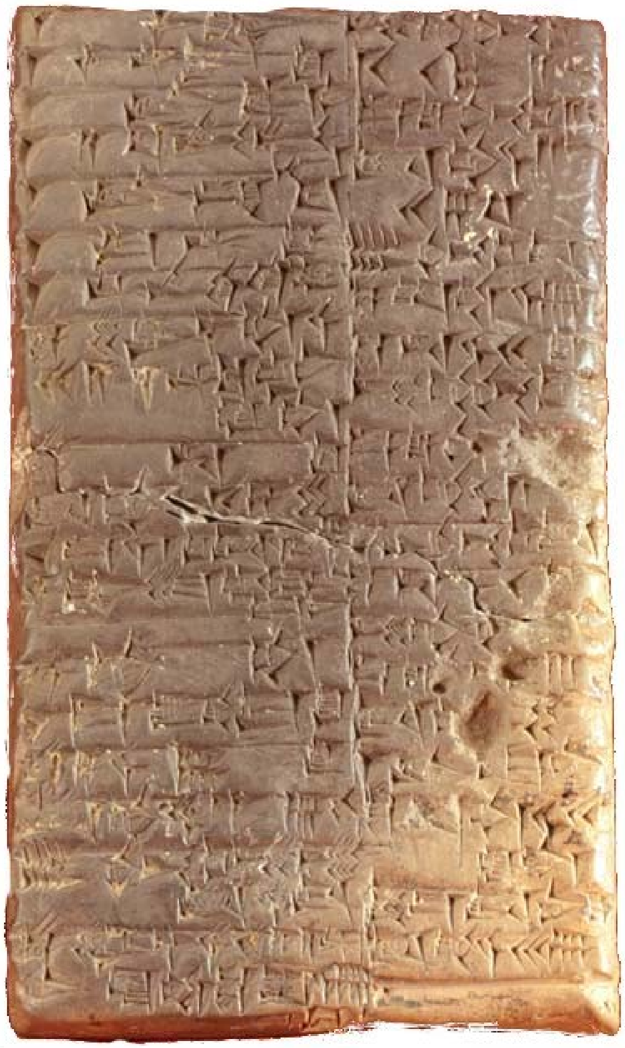
1 What is multimedia information retrieval?
- 1.1 Information retrieval
- 1.2 Multimedia
- 1.3 Semantic Gap?
- 1.4 Challenges of automated multimedia indexing

2 Basic multimedia search technologies

3 Evaluation of MIR Systems

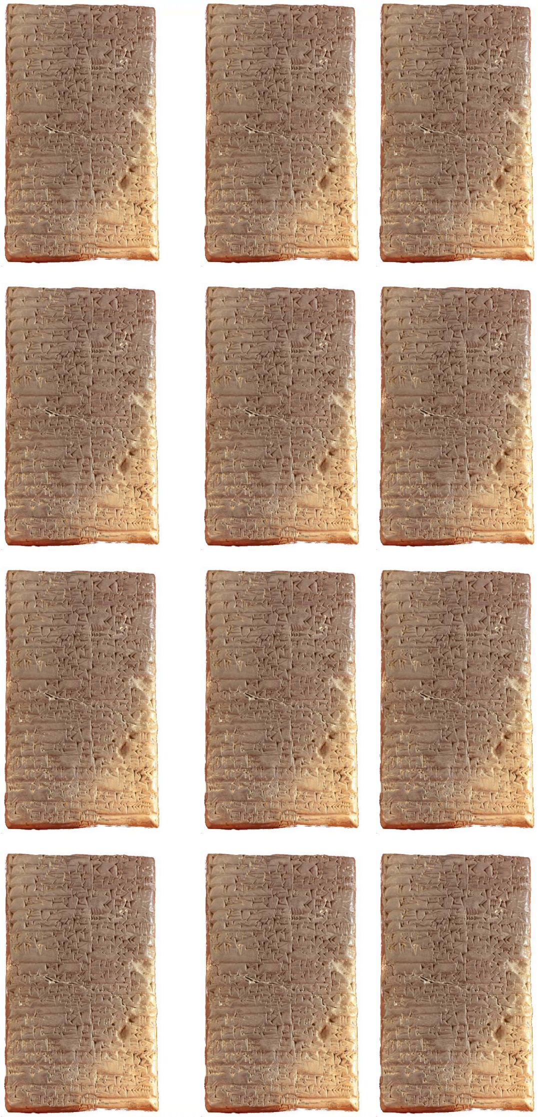
4 Added value

2400 BCE



[Kirkor Minassian collection, Library of Congress]

Incipits



1 Honored and noble warrior
2 Where are the sheep
3 Where are the wild oxen
4 And with you I did not
5 In our city
6 In former days
7 Lord of the observance of heavenly laws
8 Residence of my God
9 Gibil, Gibil [the fire god]
10 On the 30th day, the day when sleeps
11 God An [the sky god], great ruler
12 An righteous woman, who heavenly laws

[Dalby, The Sumerian Catalogs,
J library history, 21 (3), 1986]

KNOWLEDGE MEDIA
KMi
INSTITUTE

Text retrieval

368 GENERAL INDEX

The Open University

Information Retrieval

For example

“Where is the big pineapple?”

Specific (“known item”)

“Family group photo taken last Christmas”

“The song I heard at the restaurant yesterday”

General

“Family vacation pics at Surfers – like this one”

“Music to go with my vacation photo slide show”

The semantic gap

1m pixels with a spatial colour distribution

faces & vase-like object

victory, triumph, ...

disappointment, ...

What is Multimedia?

Within this lecture:

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

Multimedia queries

Built by the monks and nuns of the *Nipponzan Myohoji*, this was the first *Peace Pagoda* to be built in the western hemisphere and enshrines sacred relics of *Lord Buddha*. The Inauguration ceremony, on 21st September 1980, was presided over by the late most Venerable *Nichidattsu Fujii*, founder and …

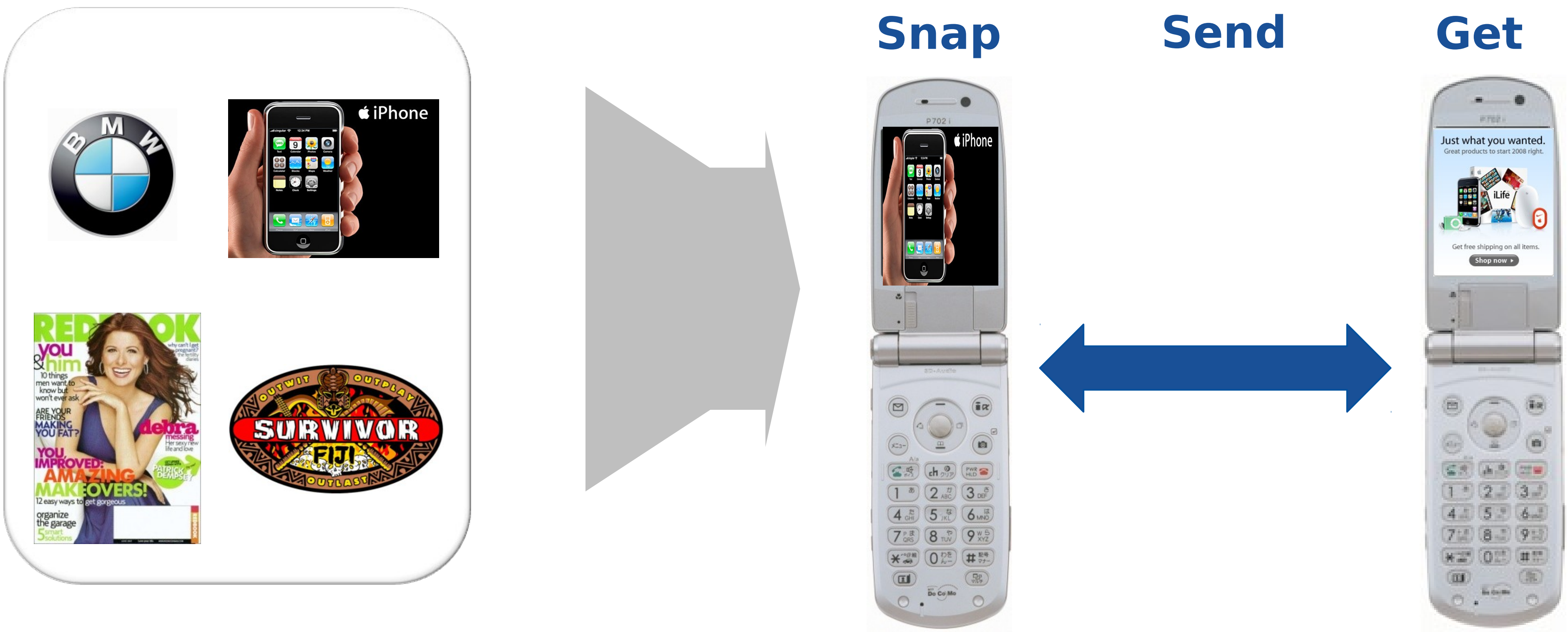
Web-based searching

"peace pagoda milton keynes"

- Google Images
- Bing Images
- Flickr
- Yahoo! Images
- ImageToss

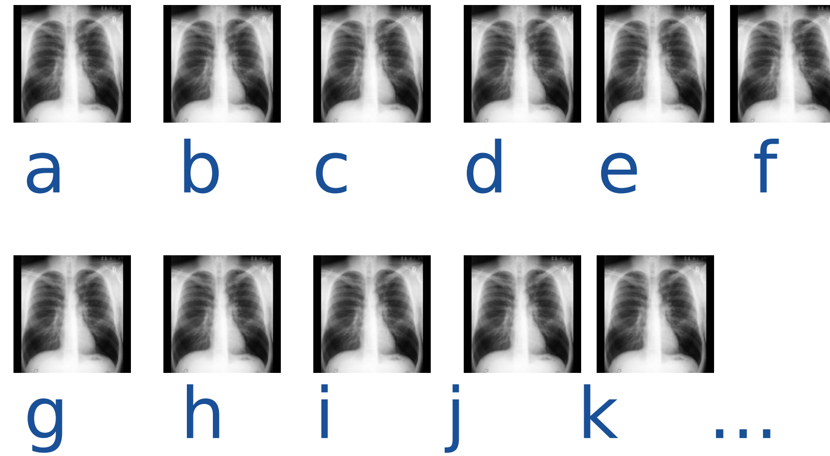
Commercial example

Snap.Send.Get™



The Open University

Medical image retrieval



[CLEF 2004 collection]

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

Exercise

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

Discuss

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

The semantic gap

1m pixels with a spatial colour distribution

faces & vase-like object

victory, triumph, ...

disappointment, ...

Polysemy

Multimedia Information Retrieval

1 What is multimedia information retrieval?

2 Basic multimedia search technologies

- 2.1 Meta-data driven retrieval
- 2.2 Piggy-back text retrieval
- 2.3 Automated annotation
- 2.4 Fingerprinting
- 2.5 Content-based retrieval
- 2.6 Implementation Issues

3 Evaluation of MIR Systems

4 Added value

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

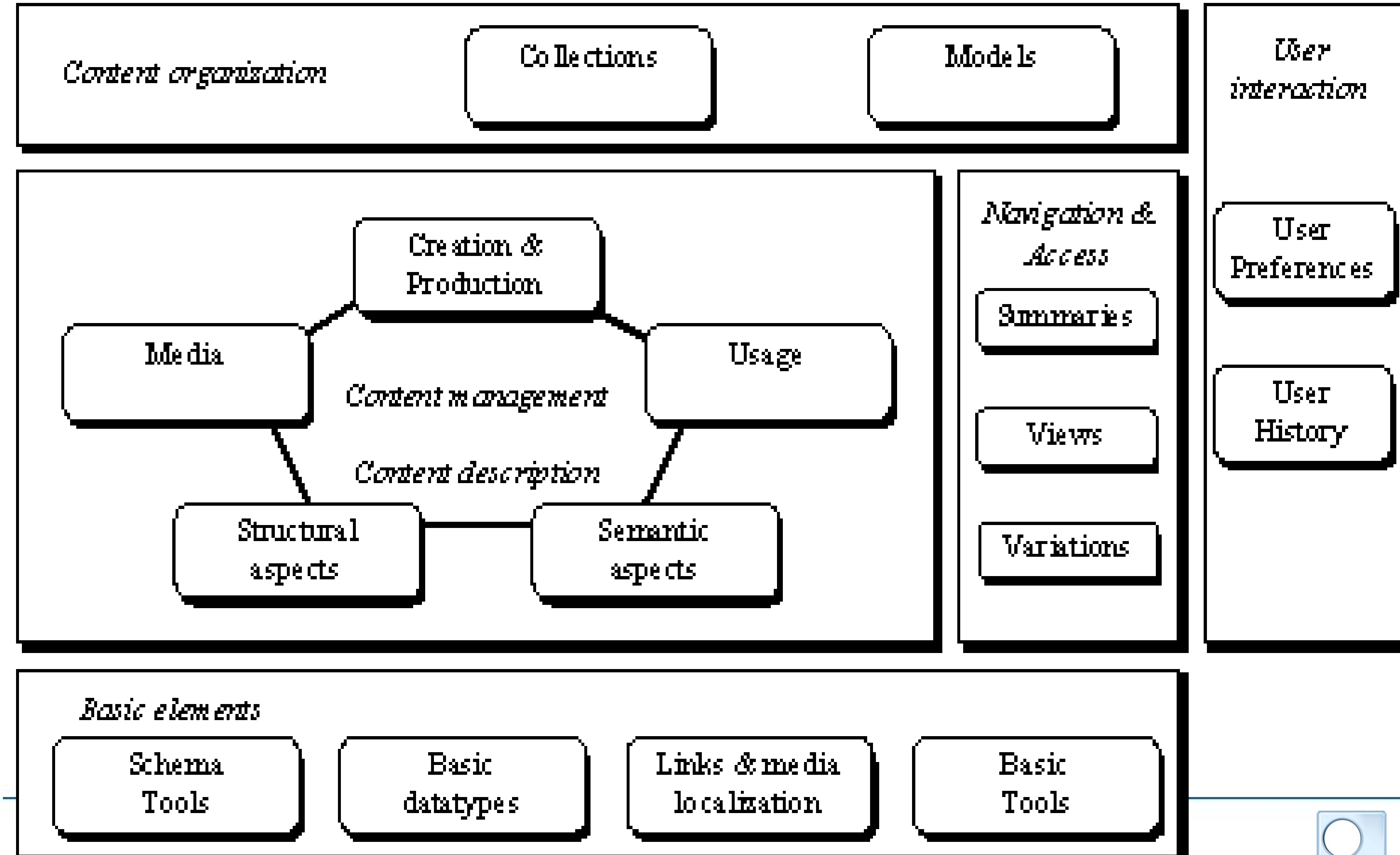
- Multimedia specific metadata standard

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



[http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm]

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>
```

The Open University

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

Metadata

Simple metadata is ambiguous (e.g. DC.creator)

DC.title = "Reconstruction of Colossus Computer"
DC.creator = "Suzanne Little"
OR
DC.creator = "Tommy Flowers"
OR
DC.creator = "Tony Sale"

Comprehensive metadata is complex
User created metadata is expensive and potentially subjective
How to create?

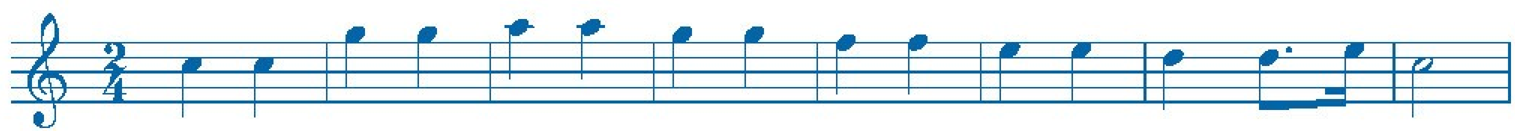
KNOWLEDGE MEDIA KMi INSTITUTE

Piggy-back retrieval

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

text

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]

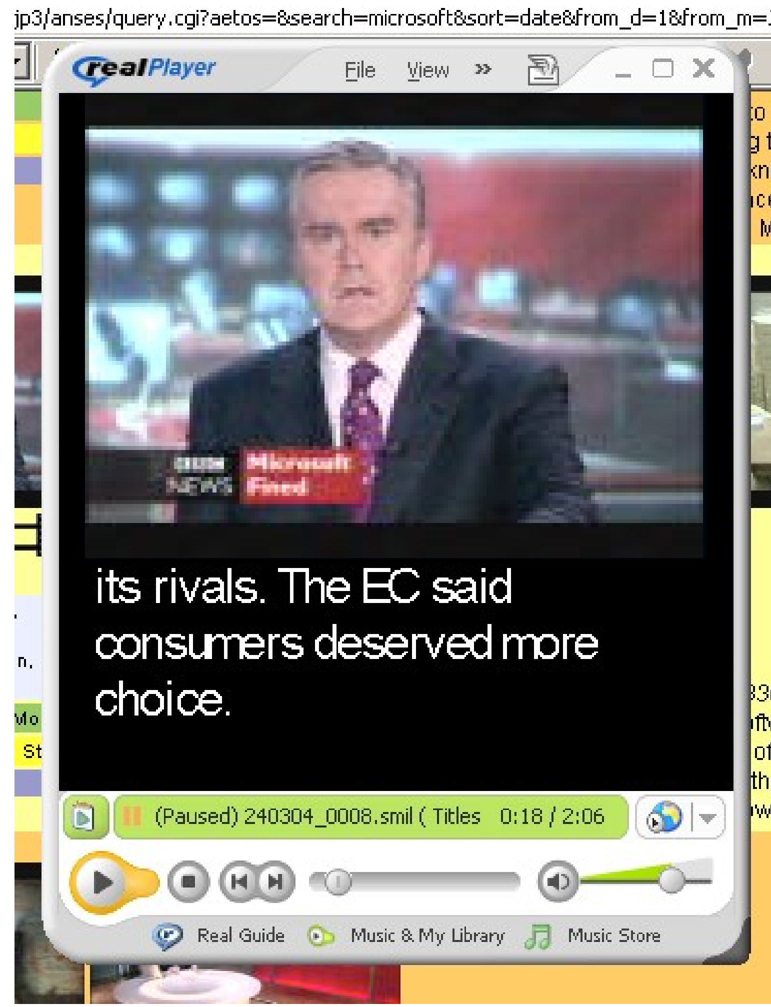
[Pickering: best UK CS student project 2000 – **national prize**]


ANSES - Automatic News Summarisation and Extraction System - Konqueror
Automatic
News
Summarization
Extraction
System
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


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.
LONDON
Play this story
Browse other news on Tue Sep 25 2007

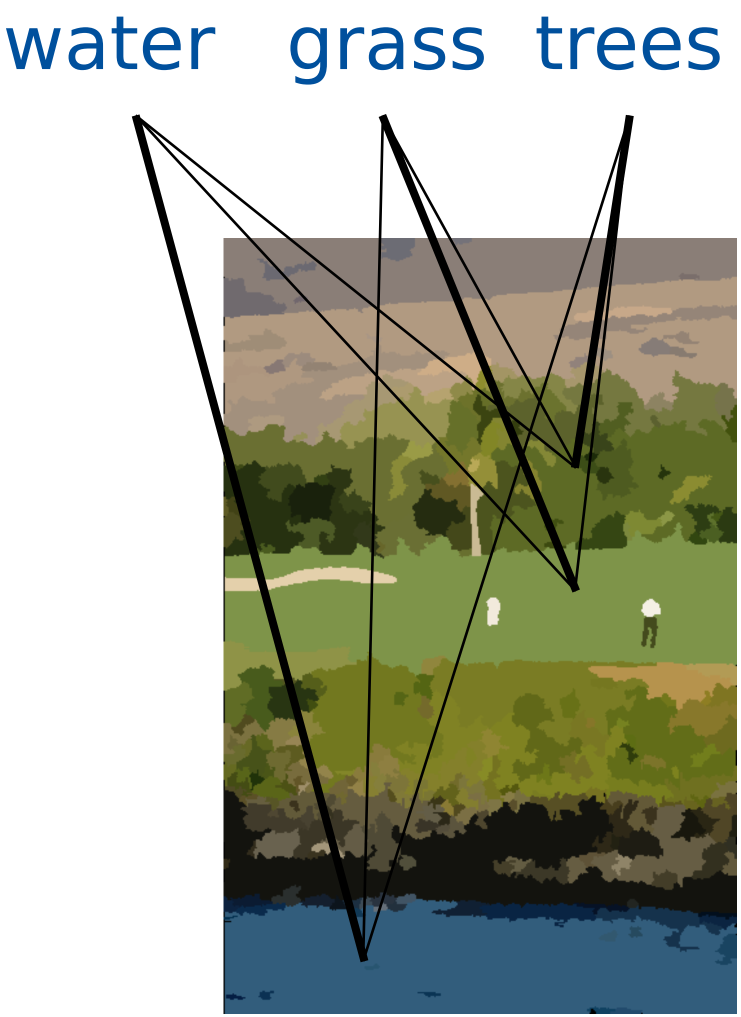
Playback with SMIL

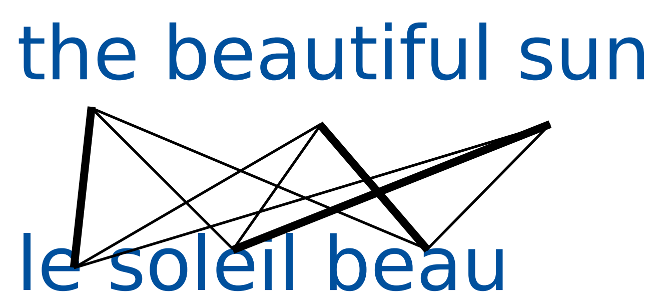
```
<smil>
<head>
<meta name="title" content="291104_0002.smil" />

<layout type="text/smil-basic-layout">
<region id = "VideoChannel" title="VideoChannel"
      left="0" top="0" height="240" width="320"
      background-color="#888888" fit="meet" />
<region id="TextChannel" title="TextChannel"
      left="0" top="240" height="120" width="320"
      background-color="#888888" fit="fill" />
</layout>
</head>

<body>
<par title="multiplexor">
<video src="content/291104_0002.rm"
      id="Video" region="VideoChannel"
      title="Video" fill="freeze" />
<textstream src="291104_0002.rt"
      id="Subtitles" region="TextChannel"
      title="Titles" fill="freeze" />
</par>
</body>
</smil>
```



KNOWLEDGE MEDIA KMi INSTITUTE

Automated annotation as machine translation





Automated annotation as machine learning

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

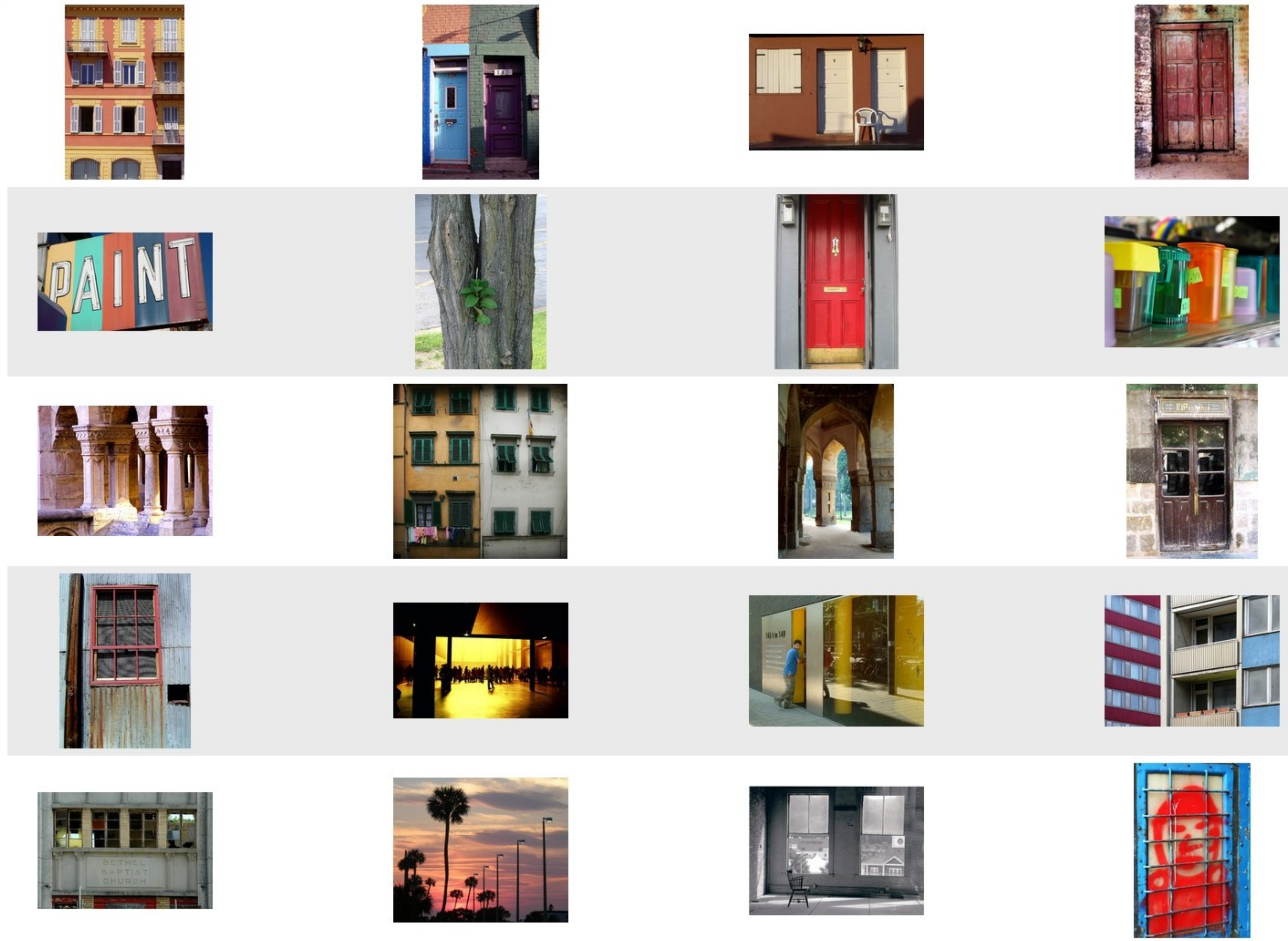
[with Yavlinsky et al CIVR 2005]
[with Yavlinsky SPIE 2007]
[with Magalhaes CIVR 2007, **best paper**]

Automated: water buildings city sunset aerial

[Corel Gallery 380,000]

The good

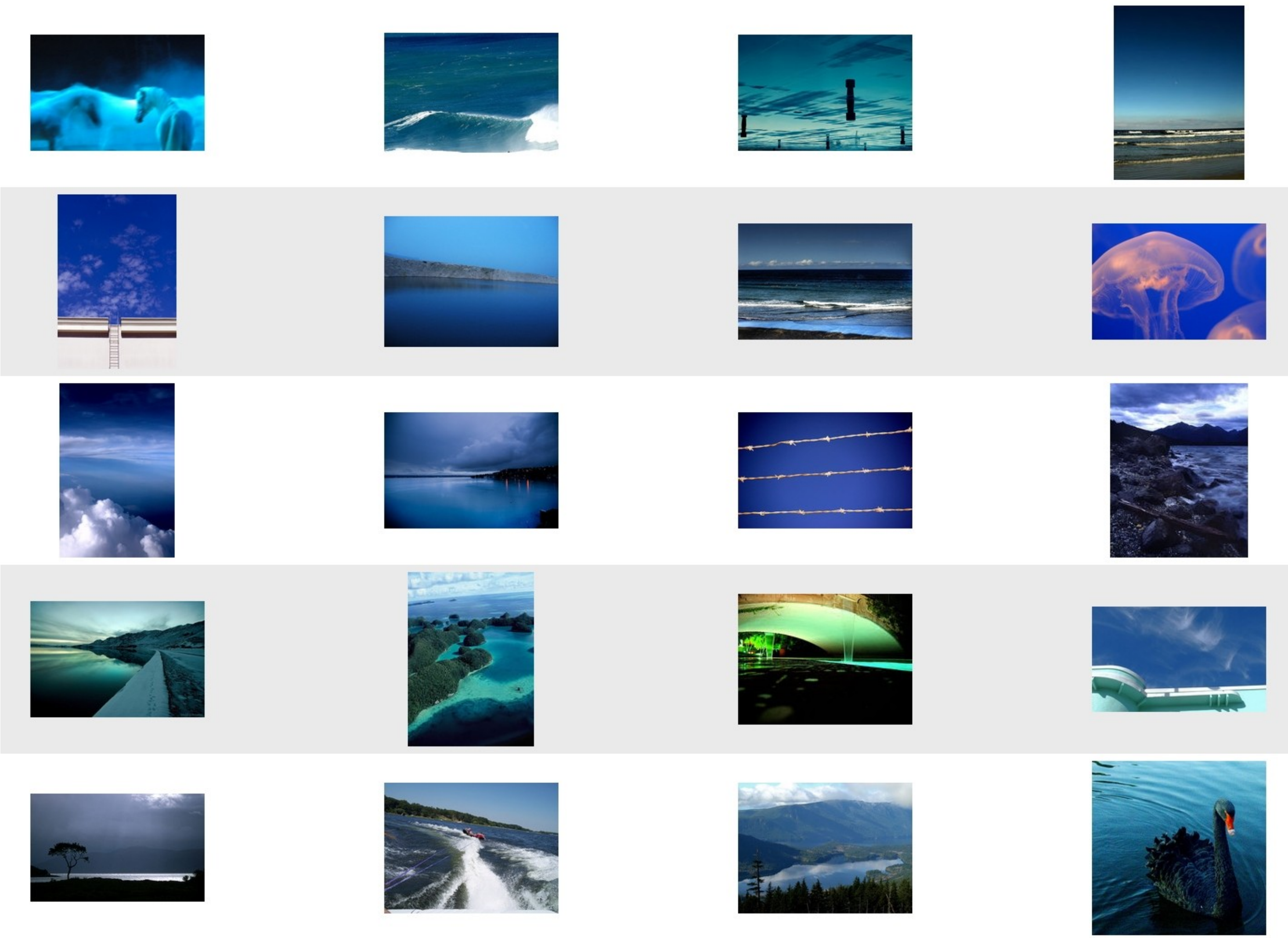
door



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

The bad

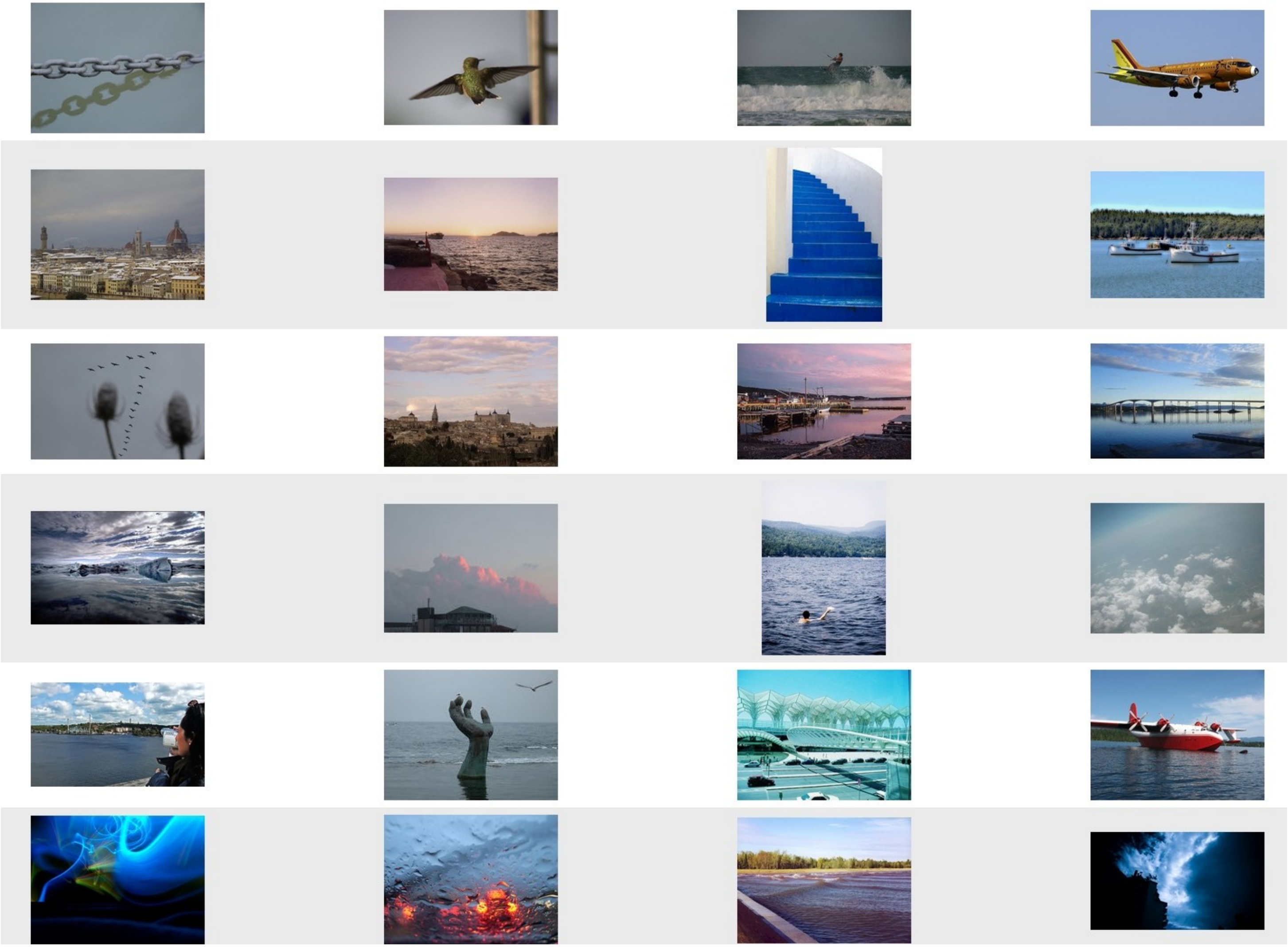
wave



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

The ugly

iceberg

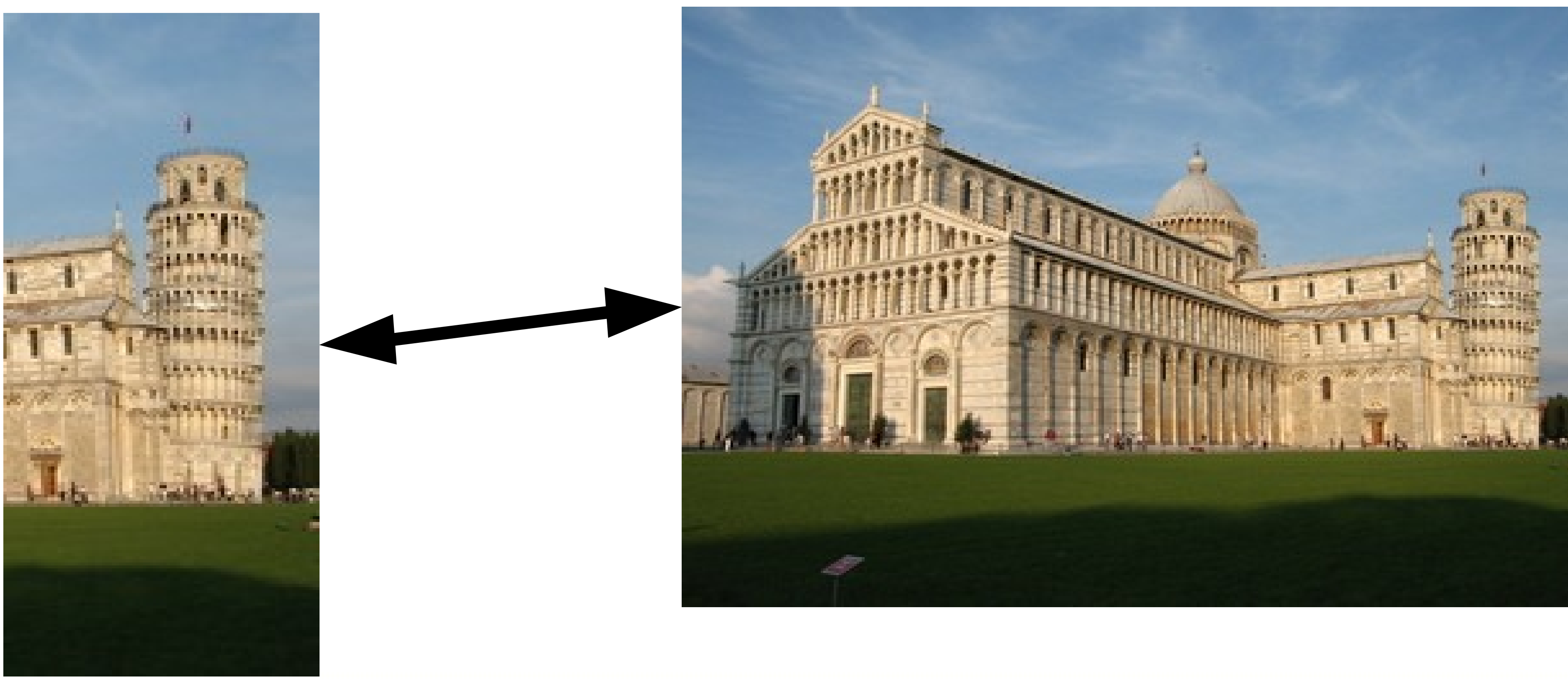


[beholdsearch.com, 19.07.2007,
now behold.cc (Yavlinksy)]
[images: Flickr creative commons]

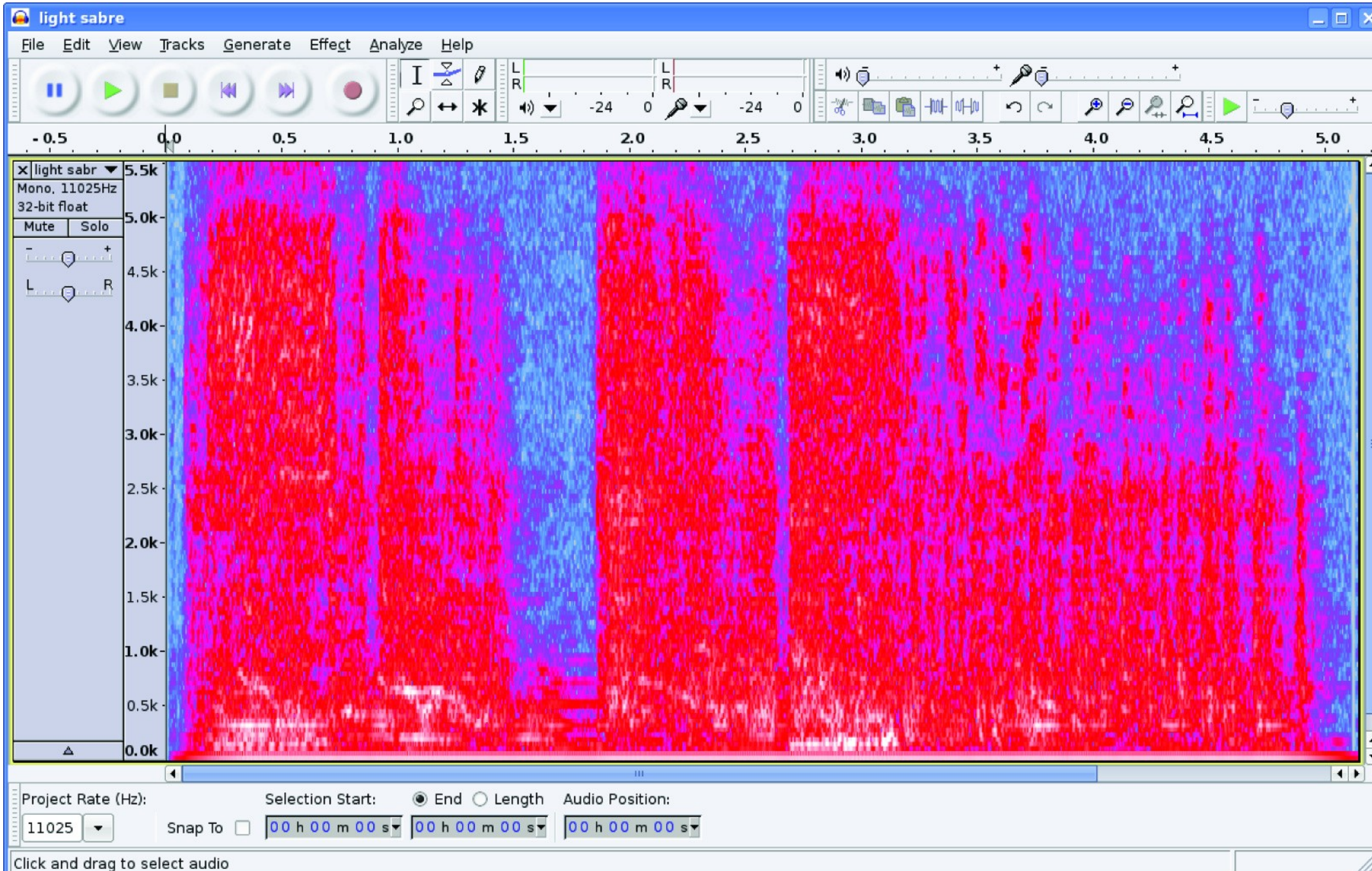
Fingerprinting

Uniquely identify multimedia objects in a database

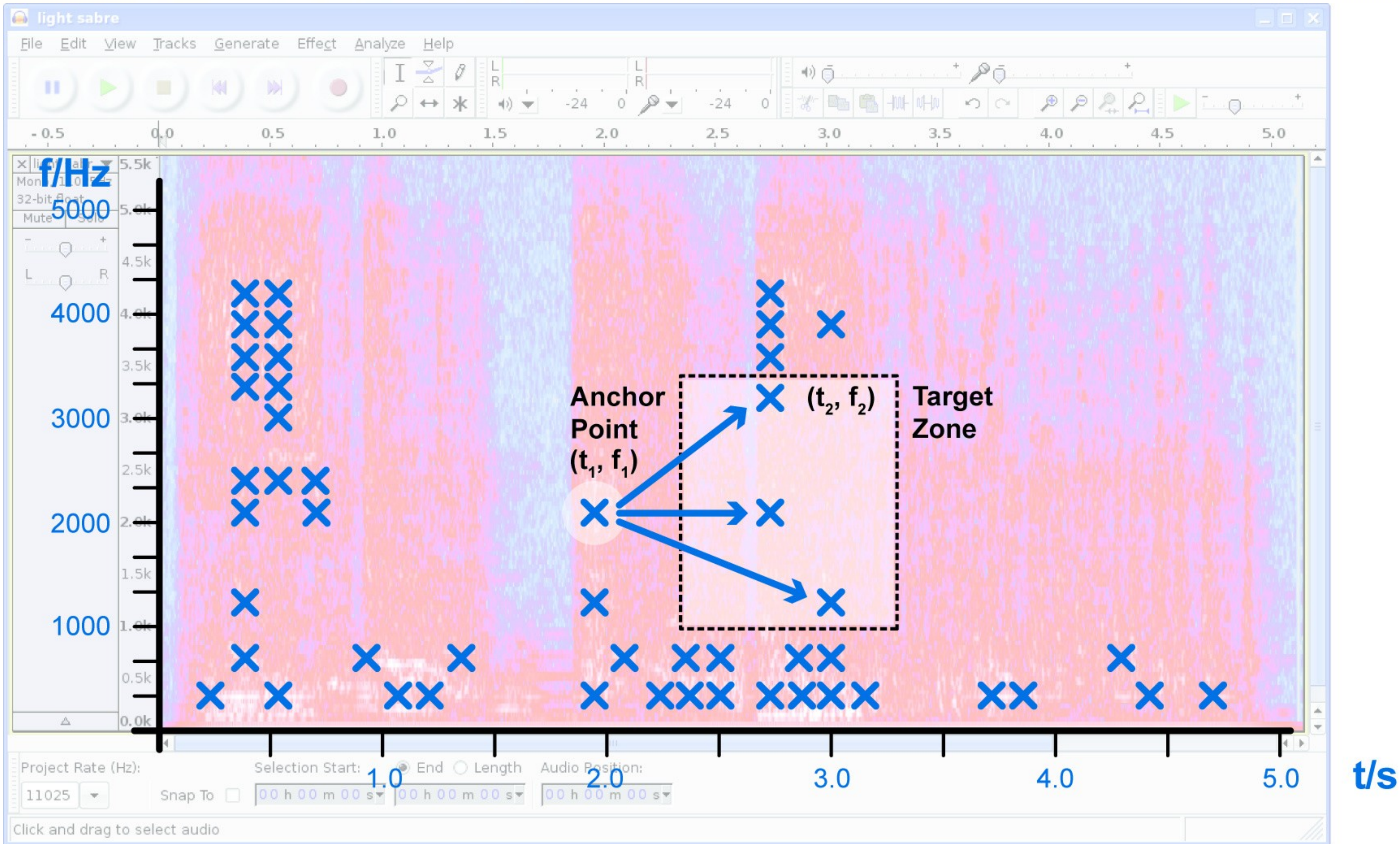
Find *specific* media based on content



Audio fingerprinting



Salient points



Encoding: (f1, f2, t2-t1)

Example applications

- Shazam [http://www.shazam.com/]
  - discover what song is playing
- Last.fm also have acoustic fingerprinting
- AudioID (Fraunhofer Institute); MusicBrainz; MusicID etc.

Image fingerprinting

$$h^i : \mathbb{R}^d \to \mathbb{Z}$$

$$v \mapsto h^i(v) = \left\lfloor \frac{a^i v + b^i}{w} \right\rfloor$$

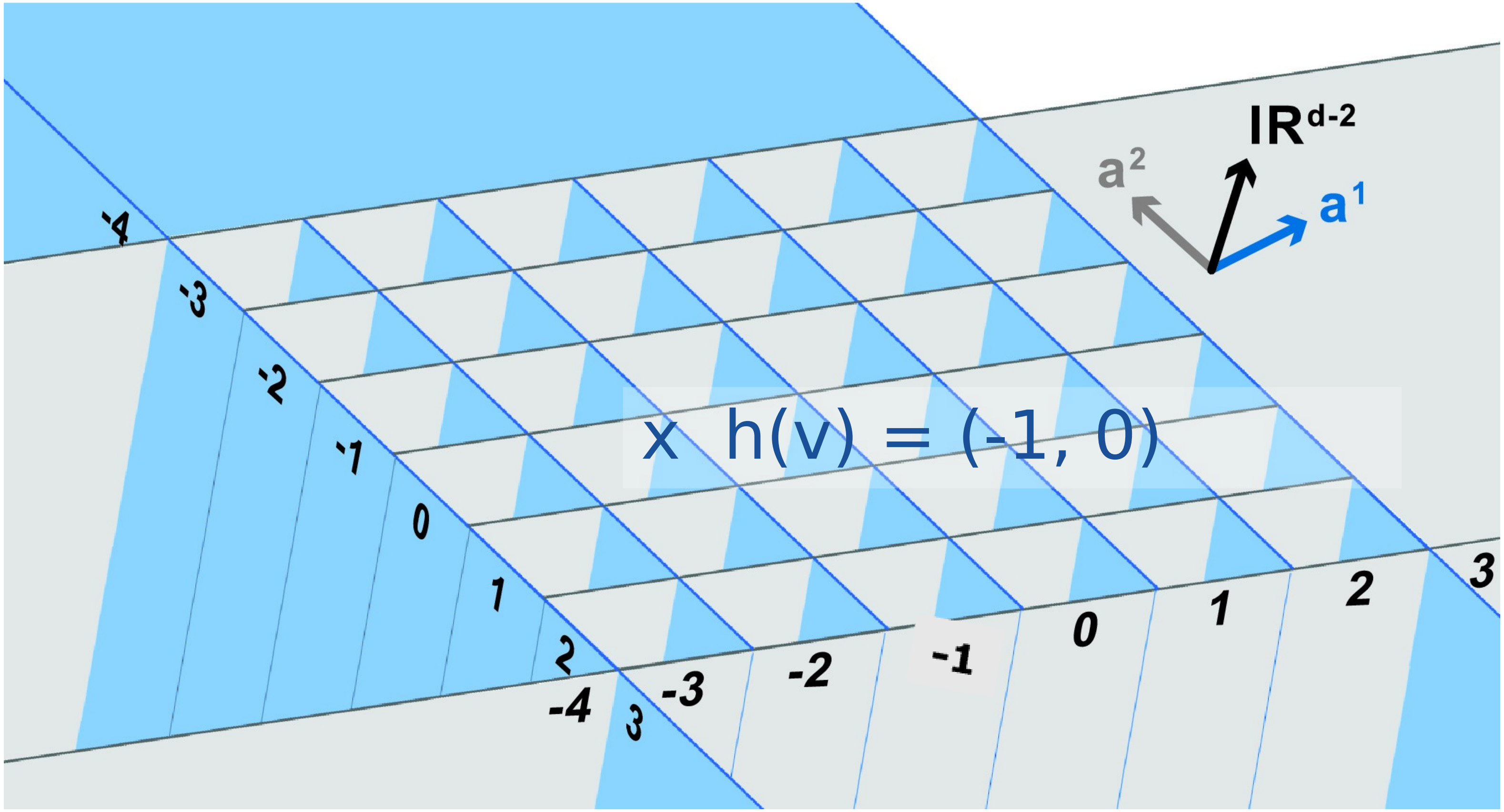
$a^i \in \mathbb{R}^d$ is a random Gaussian-distributed vector

$w \in \mathbb{R}^+$ 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.

LSH hashes



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:

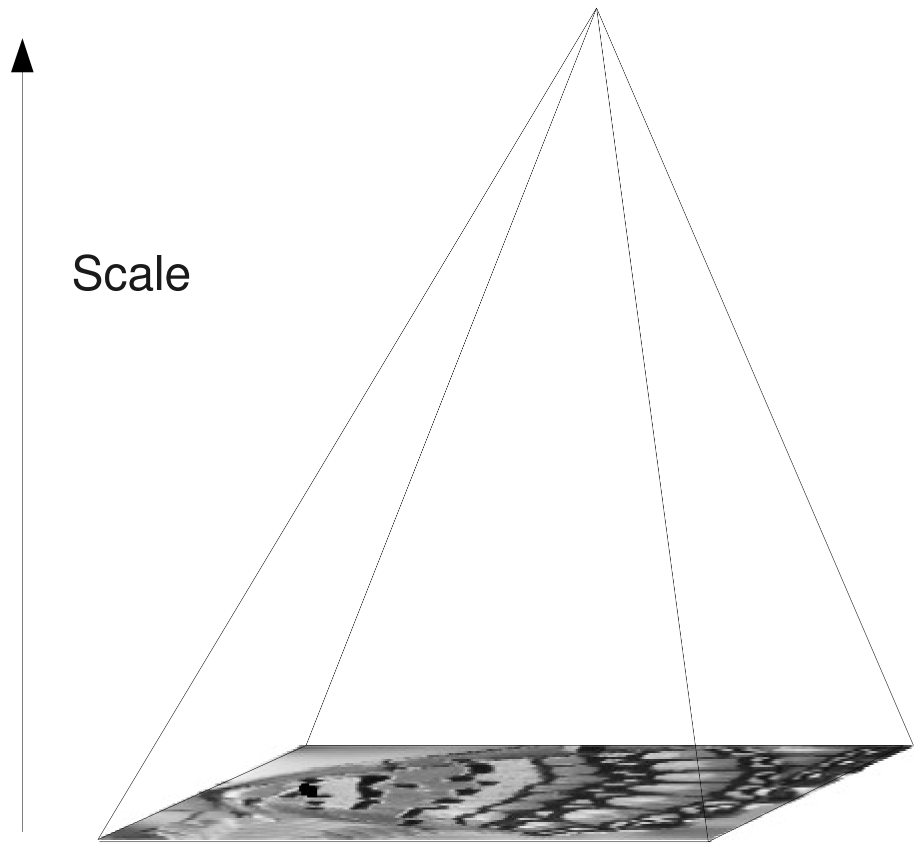
Detect scale space extrema

Localise candidate keypoints

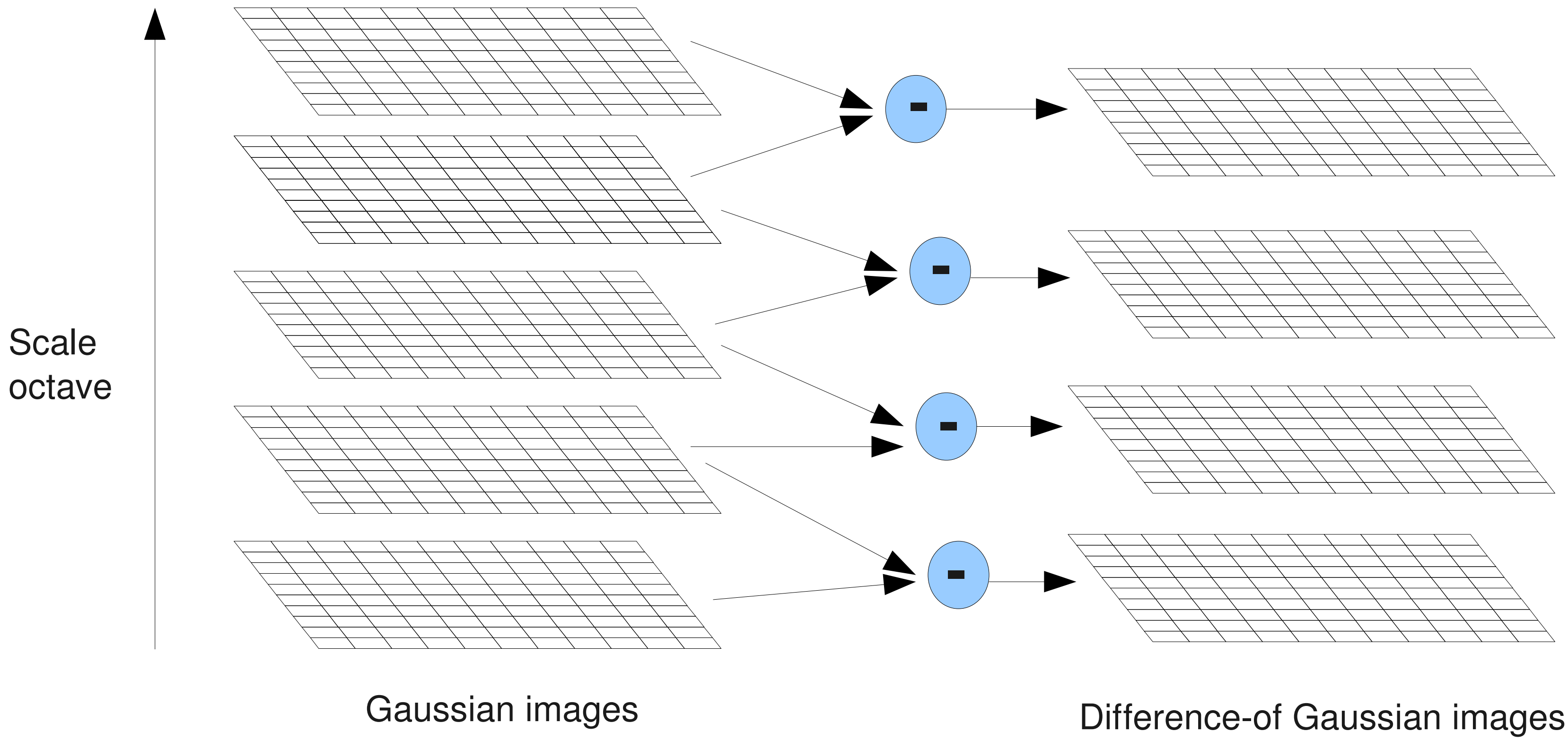
Assign an orientation to each keypoint

Produce keypoint descriptor

A scale space visualisation



Difference of Gaussian image creation



Gaussian blur illustration

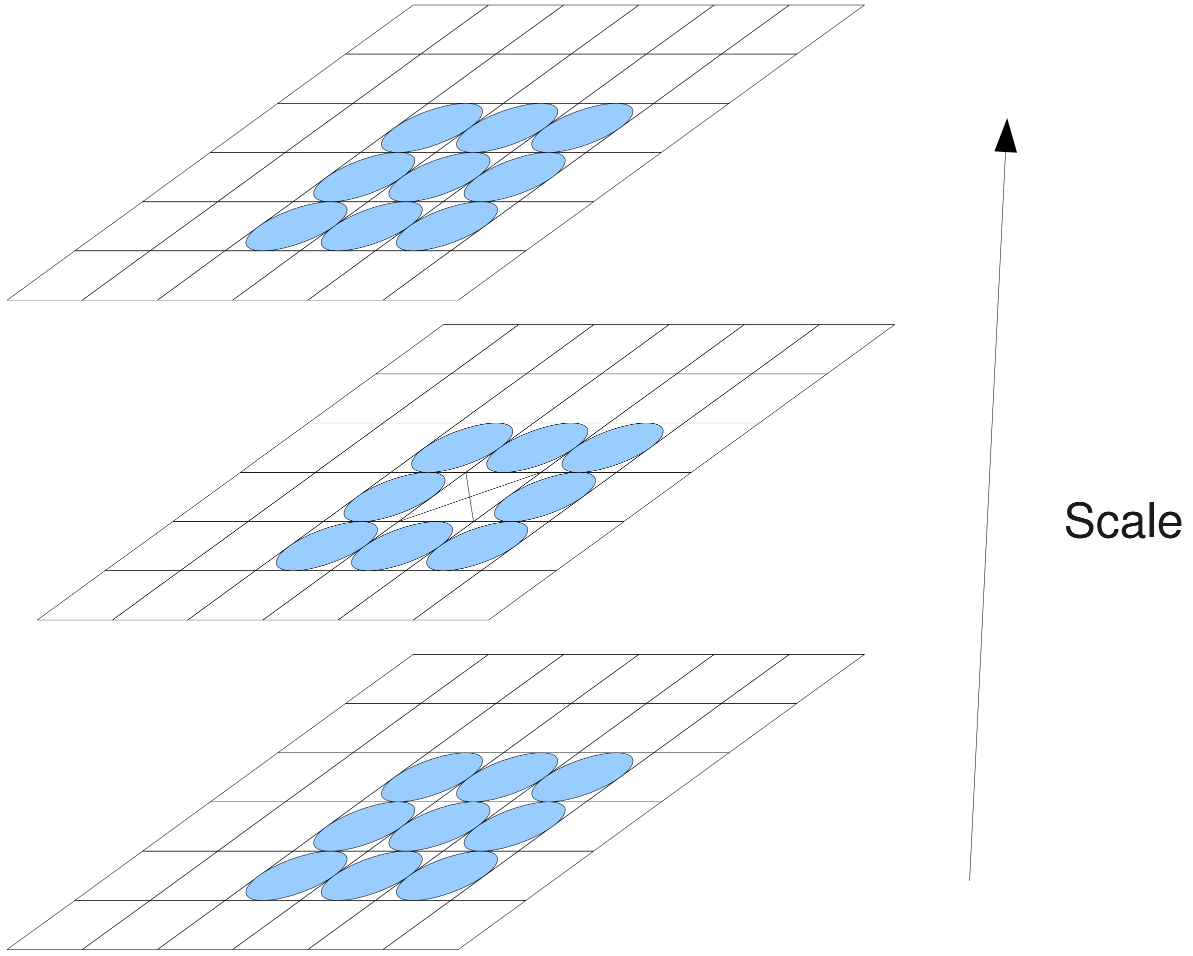
Difference of Gaussian illustration

The SIFT keypoint system

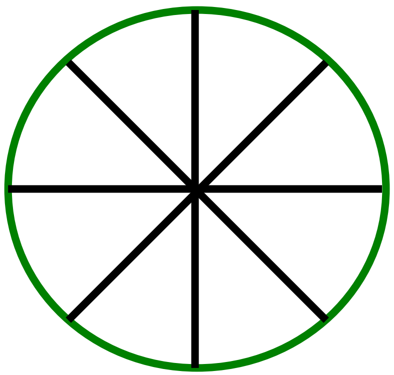
Once the Difference of Gaussian images have been generated:

- 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

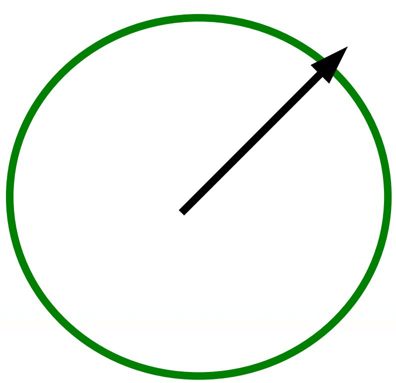


Orientation assignment

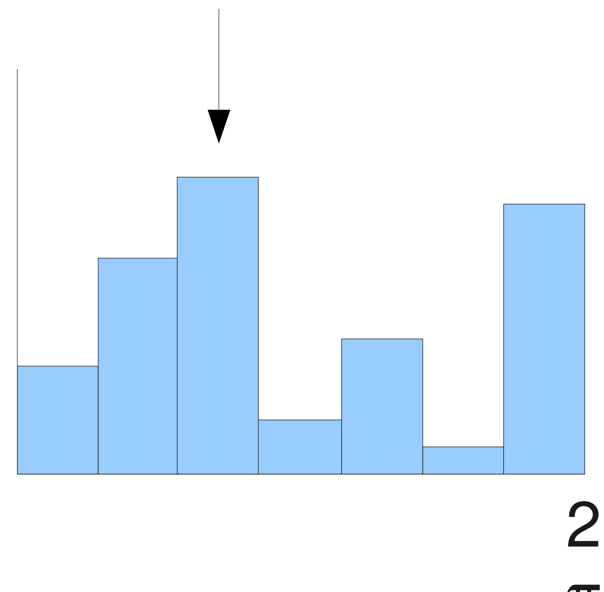


Each sample weighted by gradient magnitude and Gaussian window.

Canonical orientation at peak of Smoothed histogram.



Orientation histogram with 36 bins – one per 10 degrees.



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.

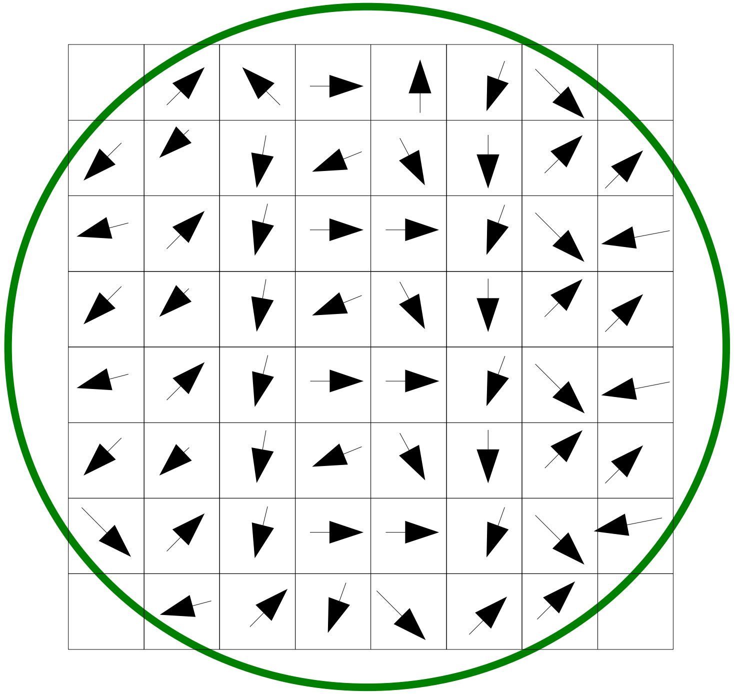
Must be as invariant as possible to changes in illumination and 3D viewpoint.

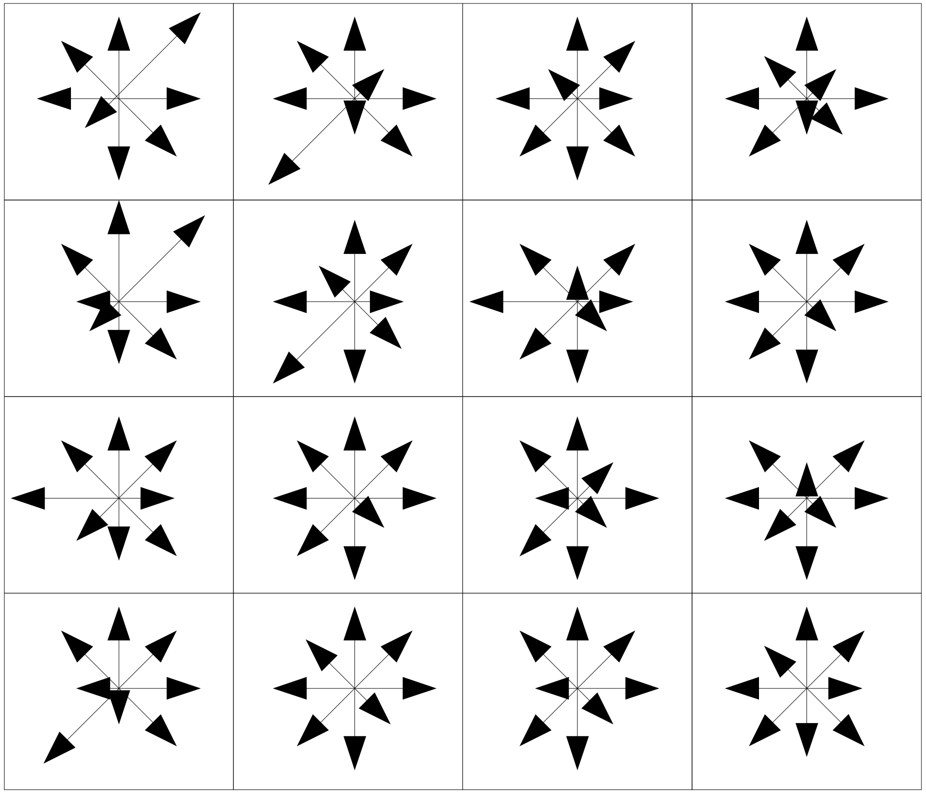
Set of orientation histograms are computed on 4x4 pixel areas.

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

SURF

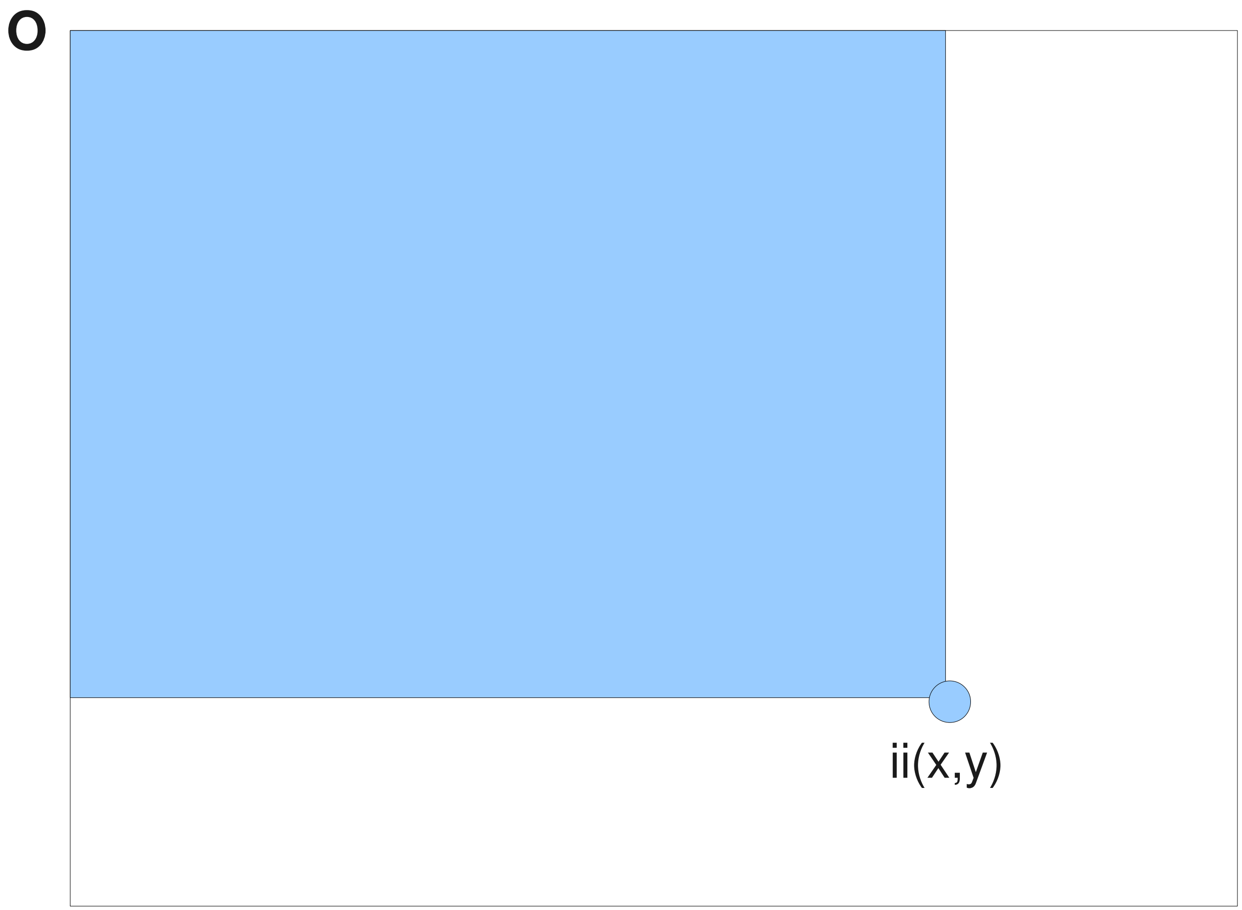
- Alternative to SIFT - “Speeded Up Robust Features”
- High dimensionality of SIFT descriptor makes it costly to compute and slow to match.
- Goal is to speed up the detection and description process for image features.
- Similar to SIFT but the authors claim better and more robust performance.

[Bay and Tuytelaars and Van Gool, H and T and L. Speeded Up Robust Features. In ECCV 2006, pp. 404-417.]

SURF

- Uses integral images (similar to summed area tables) to quickly compute box-type convolution filters.
- Integral image = the sum of the intensities of all pixels contained in the rectangle defined by the pixel of interest and the origin.

Integral image theory



The value of the integral image at point (x,y) = the sum of all pixels above and to the left.

Integral image theory

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x', y')$$

Using the following pair of recurrences:

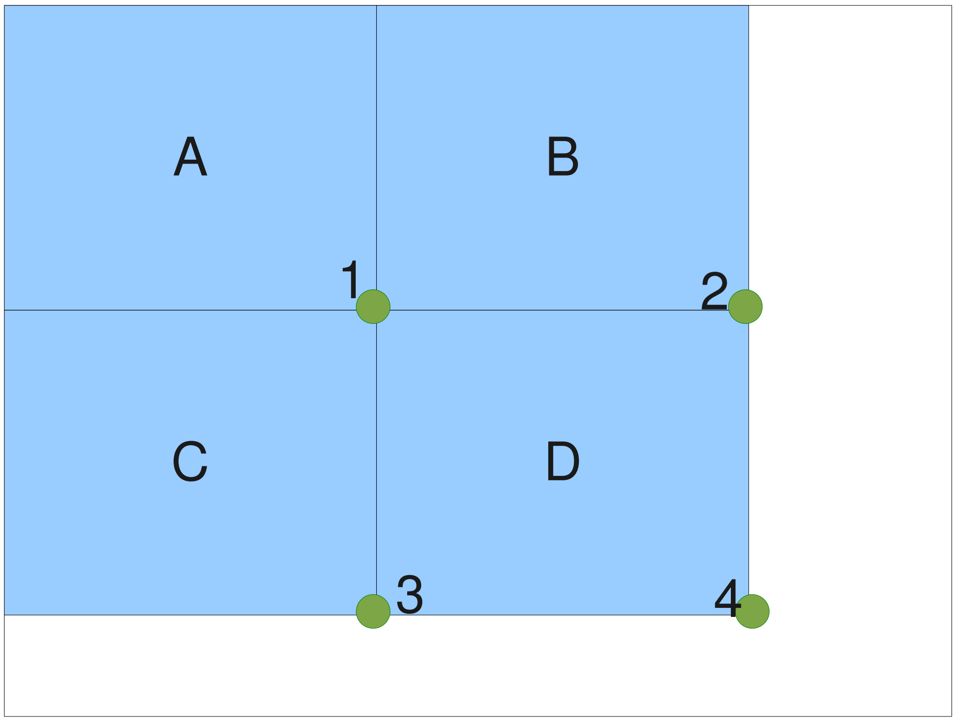
$$s(x,y) = s(x, y-1) + i(x,y)$$

$$ii(x,y) = i(x-1, y) + s(x,y)$$

Where **s(x,y)** is the cumulative row sum
s(x, -1) = 0 and
ii(-1,y) = 0

the integral image can be computed in one pass over the original image

Integral image theory

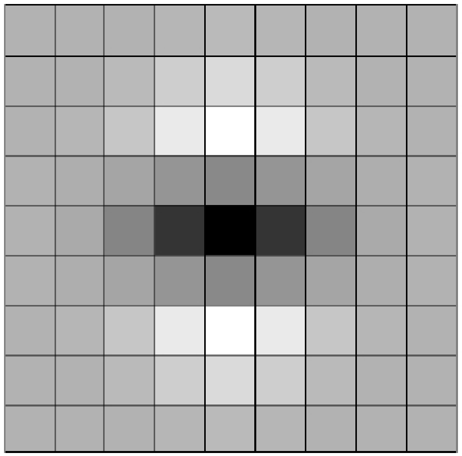


Integral image at point 1 = **sum of pixels in A.**
Value at point 2 = **A+B.**
Value at point 3 = **A+C.**
Value at point 4 = **A+B+C+D.**
Sum within D can be calculated as 4 + 1 – (2 + 3).

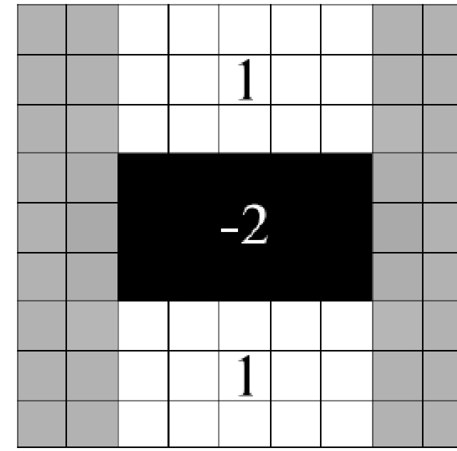
SURF detector

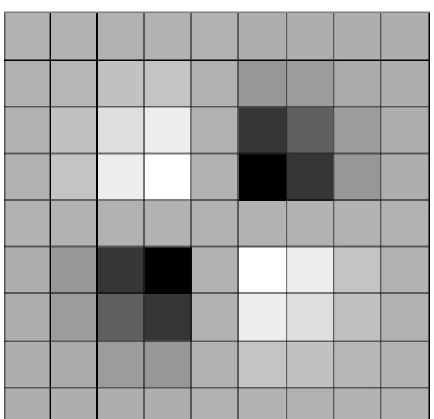
Gaussians

Box filter equivalent

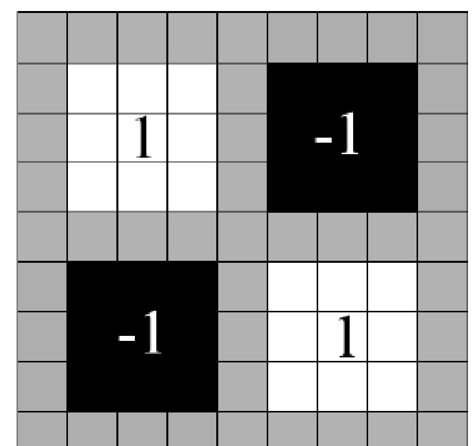


Y direction





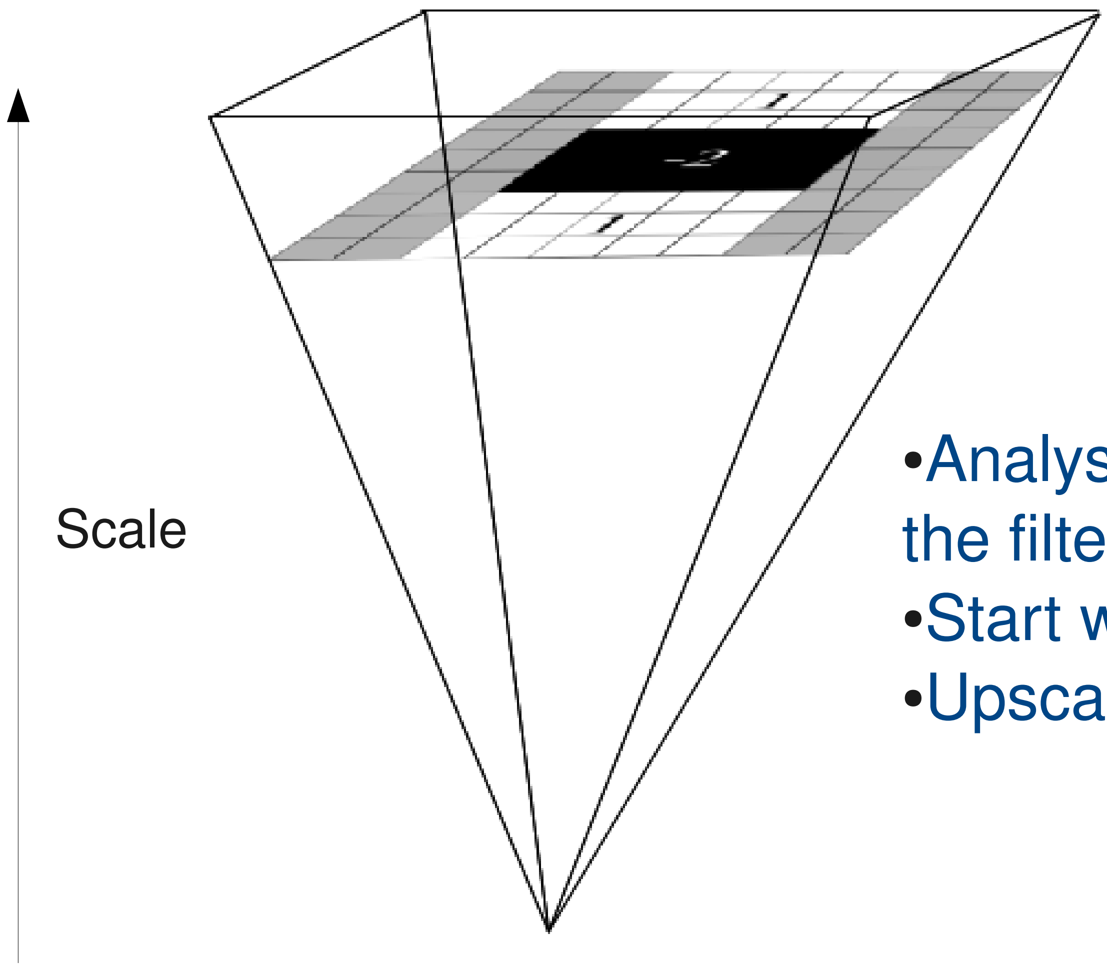
XY direction



Computation time increases with filter size.

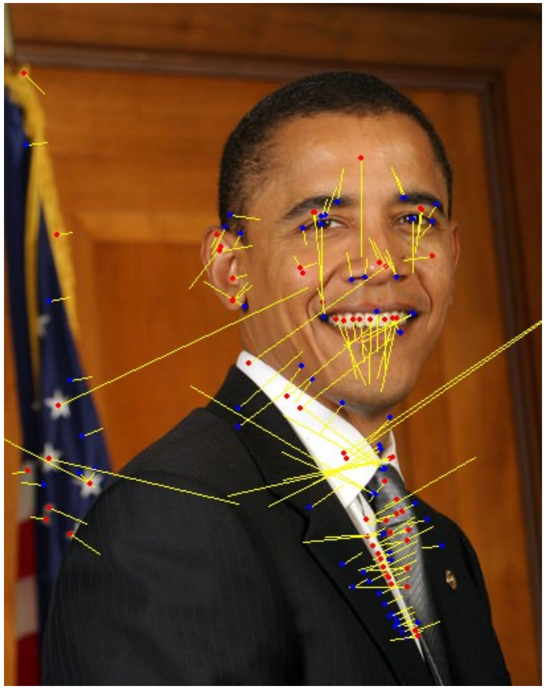
Computation time constant and Independent of filter size.

SURF Scale space



- Analyse by upscaling the filter size
- Start with 9x9
- Upscale by octaves (x2)

SURF: some example images



SURF and SIFT: differentiation

SURF and SIFT both focus on the spatial distribution of gradient information.

SURF is

- three times faster than SIFT
- less susceptible to noise (claimed to be!)
- good at handling serious image blur
- good at handling image rotation
- does not handle viewpoint change or illumination change well

SURF does not always outperform the original SIFT implementation.

1 What is multimedia information retrieval?

- 1.1 Information retrieval
- 1.2 Multimedia
- 1.3 Semantic Gap?
- 1.4 Challenges of automated multimedia indexing

2 Basic multimedia search technologies

- 2.1 Meta-data driven retrieval
- 2.2 Piggy-back text retrieval
- 2.3 Automated annotation
- 2.4 Fingerprinting
- 2.5 Content-based retrieval
- 2.6 Implementation Issues

3 Evaluation of MIR Systems

4 Added value