Multimedia searching: Techniques and systems

Dr. Nastaran FATEMI

Institute of Information and Communication Technologies, HEIG-VD, Yverdon, Switzerland

Nastaran.Fatemi@heig-vd.ch

heig-vd
Haute Ecole d'Ingénierie et de Gestion du Canton de Vaud

Objectives

- Give a brief introduction to the principles of multimedia information retrieval (MMIR)
- Present the fundamental challenges involved in building MMIR systems
- Highlight a few of the promising research directions in this field

2

Need for MMIR

- There is an ever-increasing amount of multimedia information
  - Multimedia enabled devices are emerging rapidly
  - Network bandwidth is augmenting
  - Mainstream media is moving to the Web
  - There are plenty of different applications
- Information is of no use unless you can actually access it !

3

MMIR application domains

- Medicine
  - Get diagnosis of cases with similar scans
- Law enforcement
  - Child pornography prosecution
  - Copyright infringement (music, videos, images)
  - CCTV video retrieval (car park, public spaces)
- Digital libraries
  - TV archives
  - Photo/video sharing and retrieval
  - Scientific image archives
  - Etc.

These application require searching, visualising, summarizing and browsing of MM data.

4

Text IR

- More than 30 years of research and development
- Extensive theoretical foundations
  - Retrieval models
  - Ranking algorithms
  - Indexing structures
- Effective implementations
  - Thousands of textual digital libraries and search engines, ex: Google!

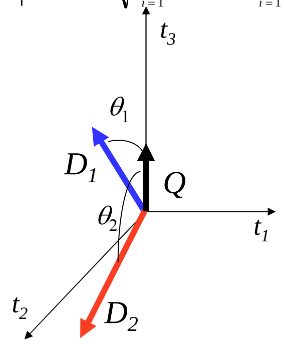
5

Text IR : principles

- Documents and queries are represented as **vectors** of term weights.
- Term **weights** represent the importance of terms within documents. They can be calculated based on the frequency of the terms in the documents (TF*IDF).
- Measuring the similarity of the document and the query vectors determines the **rank** of the documents in the result set.

CosSim($\boldsymbol{d}_j$, $\boldsymbol{q}$) =

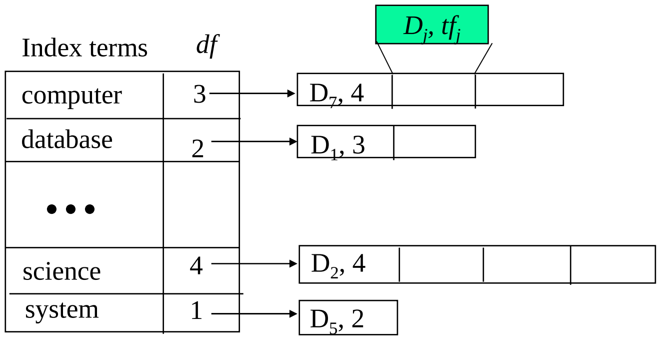
$$\frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$



6

Text IR: indexing

- Automatically built inverted indexes allow efficient resolution of queries.
- Inverted indexes give for each term a list of documents where the term appears, and the weight of the term in the document.



7

Differences between text and MM IR

- People are used to express their needs using natural language.
- Natural language queries are frequently used for text information retrieval.
- Matching between text queries and text documents is more or less straightforward.

- MM documents contain non-textual data.
- To allow text queries for MM documents, the document content should be described textually.
- MM queries are also sometimes expressed via examples or sketches.

8

Examples of MM queries

- Find pictures of rose flowers.
- Find pictures of Madrid city.
- Find video shots of people walking.
- Find video shots of at least 10 seconds showing a car race.
- Find shots with people crying victory
- Find scenes of debates between Nicolas Sarkozy and Ségolène Royal in the context of French 2007 presidential election.
- Find images similar to a given example.
- Find images similar to a given sketch.

Different categories of queries

- Queries expressed as a text description of what is desired:
  - Place/Person/Object/Event: Concepts
  - Visual/Audio/Thematic: Viewpoints
  - Shot/Scene: Granularities
- Queries expressed by providing an example/sketch similar to what is desired:
  - Low-level similarity: Colour/Shape/Texture similarity
  - High-level similarity: person, place, concept similarity

Content-based multimedia retrieval

- In the state-of-the-art, we often use the expression **"Content-based multimedia retrieval"** to indicate that the search will analyze the actual contents of the image/video.
- The term **content** in this context might refer colours, shapes, textures, or any other information that can be derived form the image/video itself.
- In this sense, content-based retrieval is opposed to **metadata based retrieval**, in which searches must rely on metadata such as keywords.

11

MM queries expressed in text

- The query is expressed in terms of "concepts".
- How to extract concepts from MM content ?
- Different approaches:
  1. Use the text parts of the MM
  2. Manually annotate concepts: provide text descriptions.
  3. Automatically extract concepts

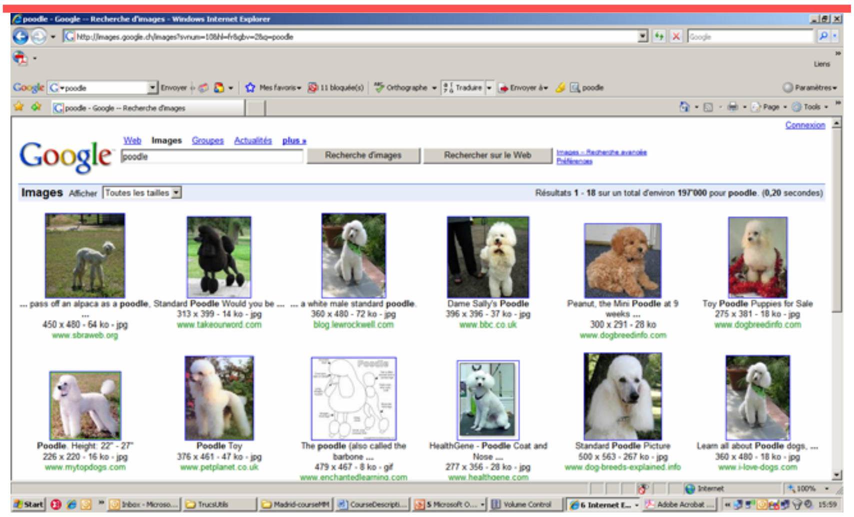
  We now present these three approaches of indexing/retrieval for Images and Videos.

12

Image retrieval

13

Image retrieval based on text



14

Image retrieval based on text (ctd.)

- Using the text surrounding the image
  - Text close to the image in the containing document
    - URL: www.healthgene.com/canine/images/poodle.jpg
    - Alt text: <img src="images/poodle.jpg" alt="Standard Poodle">
    - hyperlink: …The Standard Poodle is the original from…
- Using the text inside the image
  - Requires OCR technique
- Using manual annotations
  - Annotators (professional/non processional) can provide text descriptions of the images. This technique is quite expensive in time and resources.

Image retrieval based on text (ctd.)

- Pros
  - Easy to implement and use
  - Useful for simple and non-professional image retrieval
- Cons
  - It is incomplete and subjective
  - Some features are difficult to define in text such as texture or object shape
  - It is difficult to describe all image contents

Image retrieval based on text (ctd.)

Old or young woman?

One or two faces?

- *An image is worth 1000 words!*
- However, these 1000 words may differ from one individual to another depending on their perspective and/or knowledge of the image context.

17

Content-based image retrieval (CBIR)

- The commonly accepted way is to show a sample image, or draw a sketch of the desired images to the system, and ask the system to retrieve all the images similar to that sample image or sketch.
- Examples
  - retrievr
  - Imageseek

http://labs.systemone.at/retrievr

18

CBIR (Ctd.)

- CBIR relies on visual features that can be **automatically extracted** from the image. These features include:
  - Low/pixel level features describing the colour, texture, and/or (primitive) shapes within the image.
  - The objects, identified within an image.
- Visual descriptors are used to form the basis for one or more image **signatures** that can be indexed.
  - An image query is analyzed using the same descriptor techniques giving a **query signature**, which is then compared to the image signatures to determine similarity between the query specification and the database image signatures.

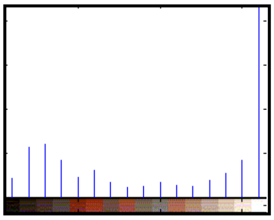
19

Images containing similar colours

- Examining images based on the colours they contain is one of the most widely used techniques because it does not depend on image size or orientation.
- Various techniques are used:
  - Colour histograms
  - Colour Moments
  - Colour Sets: Map RGB Colour space to Hue Saturation Value, & quantize
  - Colour layout: local colour features by dividing image into regions
  - Colour correlograms

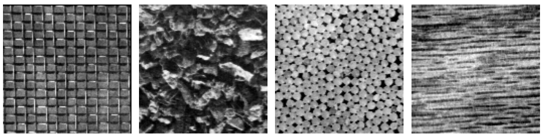
20

Colour histograms



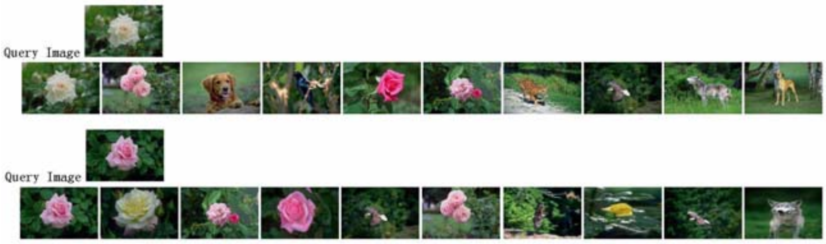
- Colour histograms give an estimation of the distribution of the colours in the image.
- The colour space is partitioned and for each partition the pixels within its range are counted, resulting in a representation of the relative frequencies of the occurring colours.

Images containing similar texture



- Texture measures look for visual patterns in images and how they are spatially defined.
- Various techniques are use
  - Co-occurrence matrix
  - Orientation and distance on gray-scale pixels
  - Wavelet Transforms
  - Gabor Filters
  - Tamura features corresponding to human visual perception (*coarseness*, *contrast*, *directionality*, *linelikeness*, *regularity*, *roughness)*

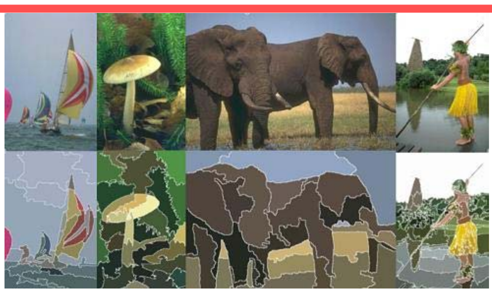
Images containing similar shape



- Various techniques are used:
  - Outer Boundary based vs. region based
  - Fourier descriptors
  - Moment invariants
  - 3-D object representations using similar invariant features
  - Well-known edge detection algorithms.

23

Images containing similar shape (ctd.)



- Shapes are often determined by first applying **segmentation** or **edge detection** to an image.
- In some cases accurate shape detection requires human intervention because methods like segmentation are very difficult to completely automate.

24

Images containing similar content

- Challenge
  - The term “similarity” has different meaning for different people.
  - Even the same person uses different similarity measures in different situations.
  - Similarity of the “content” is hardly measurable using low-level features!

25

Variety of Similarity

Degree of difficulty

- Similar colour distribution ← Histogram matching
- Similar texture pattern ← Texture analysis
- Similar shape ← Image Segmentation, Pattern recognition
- Similar real content ← Eternal goal :-)

26

The semantic gap

- Computers are very good at automatically computing **low-level features** such as colour histograms.
- Computers are rather poor in extracting **high-level semantic features**, such as objects and their meanings, actions, feelings, etc.
- High level features are indeed more useful than low-level features in content-based retrieval!

27

The semantic gap

- The gap between low-level features and high-level semantic descriptions is often referred to as the **semantic gap**.

28

Bridging the semantic gap

- The low-level features are practical from a computational point of view.
  - If a system uses these low-level features, it has to provide a way to **bridge the semantic gap** between these features and the high-level semantics required by the users.
- Establishing this bridge has turned out to be very difficult.
  - All steps towards reducing the semantic gap represent a significant leap from the current state-of-the-art in content-based retrieval.

29

Bridging the semantic gap

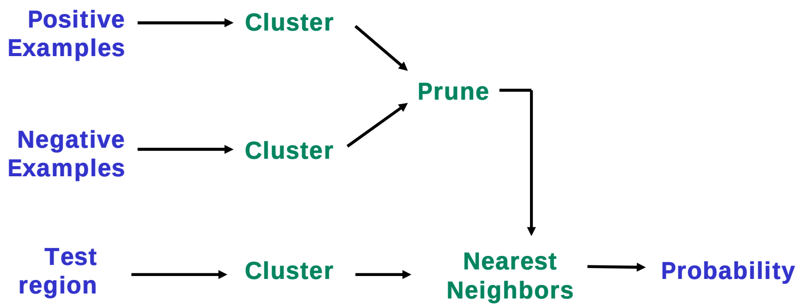
- One approach is to use **machine learning techniques** to combine different low-level features into semantic concept models.
- Example

  (Taken from Stefan Rüger et al., Content-based Multimedia Information Retrieval: Challenges & Opportunities)
  - Region segmentation + region classification (grass, water, etc.)
  - Using simple models for complex concepts (grass + plates + people = barbecue)

30

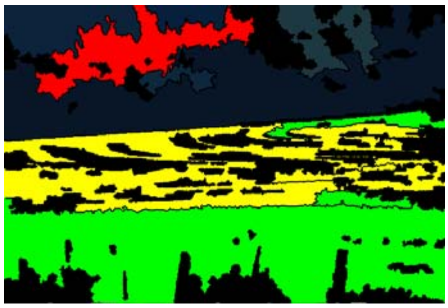
Region classifiers

- Visual categories
  - grass, sky (blue), sky (cloudy), skin, trees, wood, water, sand, brick, snow, tarmac
- Give regions a probability of membership



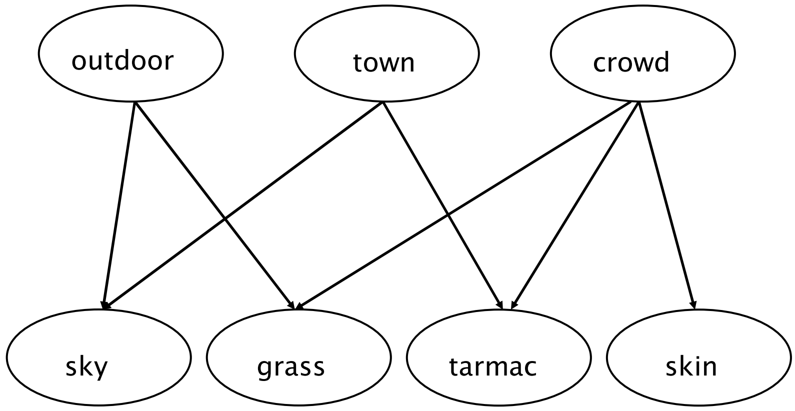
Example: grass classifier

very likely
may be
probably not



Modelling semantic concepts

Bayesian networks



33

Bridging the semantic gap

- Another approach is using **relevance feedback** in which the system learns from the user:
  - The user provides a first query and asks for similar images;
  - The system returns a set of images considered as similar to the user's query;
  - The user re-ranks the results based on his/her notion of similarity;
  - The system re-computes optimal parameters for this specific query automatically.

34

CBIR: discussion

- Problems
  - One must have an example image.
  - Example image is 2-D, hence only that view of the object will be returned.
  - Large amount of image data is required.
  - Similar colour/texture/shape does not equal similar image.
- Compromise
  - Usually the best results come from a combination of both text and content searching
    - Use existing texts (title, subject, caption)
    - Use content information (colour, texture, shape, etc.)

Video retrieval

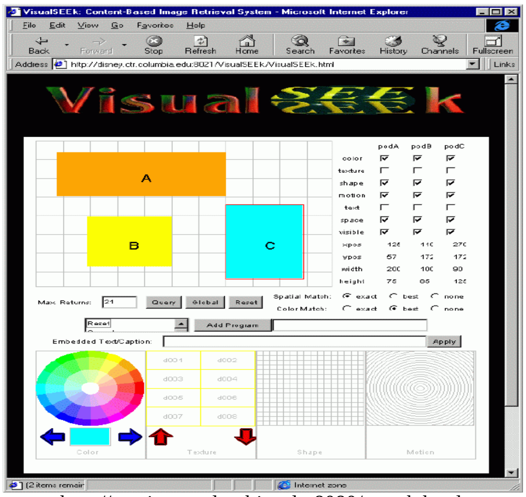
Video retrieval approaches

- Text-based retrieval
  - Same as for images, videos can be described by text and text-based indexing/retrieval methods can be applied to retrieve videos.
- Content-based retrieval
  - Video is a spatio-temporal media
    - Spatial features: ex. shape of the objects in the frames
    - Temporal features: ex. movement of objects, temporal relationships between events, duration of shots, etc.
  - Content-based retrieval should allow spatio-temporal content description
    - Ex. retrieve a sunset (yellow circle going from the top to the bottom
    - Ex. retrieve a scene showing a car race with the duration of at least 10 seconds

37

Content-based video retrieval

- VisualSEEK (Columbia University)



http://persia.ee.columbia.edu:8080/search.html

38

Text-based video retrieval

- Query by example tools have a limited scope of application.
- In practice, most of the video retrieval systems, such as TV and film archives require high-level semantic querying and browsing tools.
- Such tools require exploiting the video logical structure and the video high-level semantic descriptions.

Video content description

- Automatic description
  - **Text parts** of video, such as subtitles and Tele-texts are automatically detected using OCR techniques.
  - **Speech parts** of the video are automatically recognised (ASR) and transcribed. The text is used to index the video.
  - **Concepts** such as persons, faces, objects, are automatically detected and used for text description.

Video content description (ctd.)

- Manual description
  - Human annotators (professional/non professional) manually attach descriptions to video parts.
  - There are still many archive systems which use this approach.
- Semi-automatic description
  - Totally automatic description not yet realistic.
  - Manual description is expensive.
  - In practice, most systems use a mixture of automatic and manual descriptions.

41

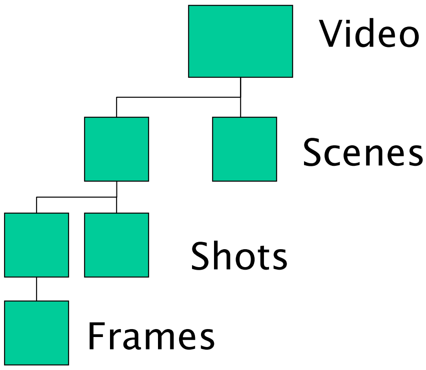
Video segmentation

- Video descriptions are usually attached to parts of video and not the whole video.
- These parts correspond to the structure of video.
- If dealing with text, then text structure is obvious:
  - paragraph, section, topic, page, etc.
  - all text-based indexing, retrieval, linking, etc. builds upon this structure;
- If dealing with video, then first it needs to be structured, automatically.

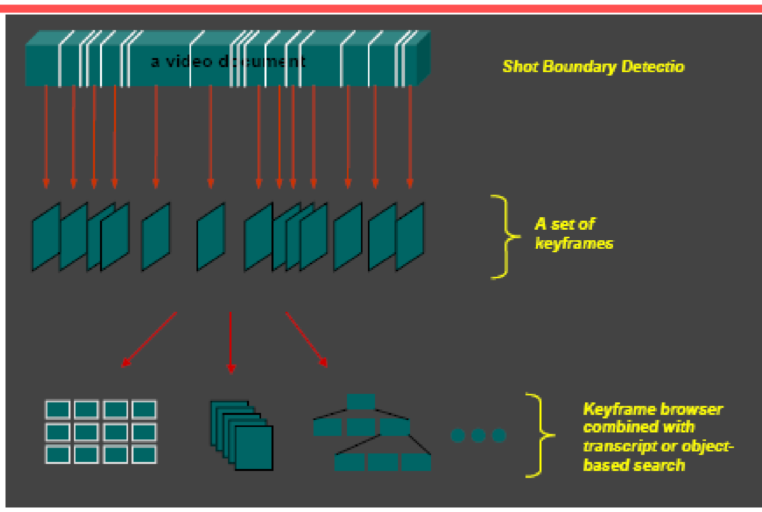
42

Automatic structuring of video

- Video “programmes” are structured into logical scenes, and physical shots.



Automatic structuring of video (ctd.)



Video temporal segmentation

- **Shot** is a sequence of frames generated during a continuous camera operation.
  - Shots are separated by different types of transitions (cuts, progressive transitions such as fade-in, fade-out, dissolves or wipes)
- There are various algorithms for **automatic shot transition detection**
  - Colour histogram comparison
  - Edge detection (good for detecting gradual transitions)
  - MPEG macroblockes
  - Combination of different approaches

45

Video temporal segmentation (ctd.)

- **Scene** is collection of usually adjacent shots focusing on the same objects or describing the same actions that are related in time and space.
- **Scene detection** is more complicated than shot detection, as it requires a higher level of semantic analysis of audiovisual content.
  - a first step of shot boundary detection,
  - followed by applying various methods to regroup shots into scenes.

46

Keyframe selection

- **Keyframes** are representative frames of a video segment.
- Keyframes are very useful
  - they are suitable for video indexing, browsing and retrieval.
  - they reduce video down to manageable size
- There are various algorithms for **keyframe detection**.
- The keyframe should be selected in a way that its visual content is the best representative of the corresponding segment.

47

Video content description

- Video segments are described based on different features
  - Camera movements
    - Camera movement detection algorithms
  - Objects/persons present in the segment
    - Object shape/movement detection
    - Face detection
    - Face recognition
  - Automatic text descriptions
  - Manual text descriptions

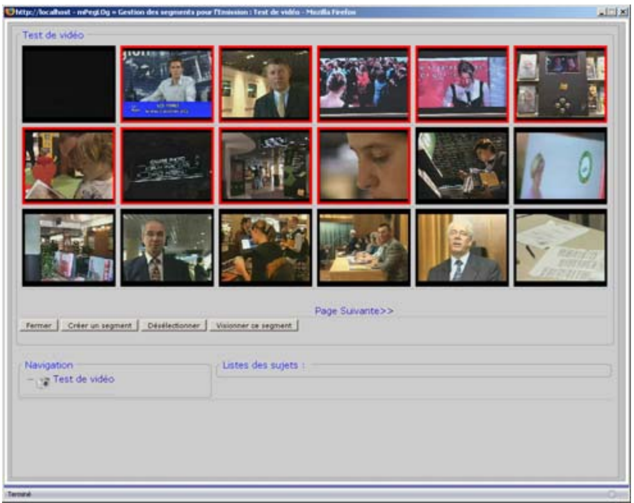
48

Video content description (ctd.)

- Example: MpegLog developed at HEIG-VD (University of Applied Sciences of Western Switzerland, Yverdon).
  - Semi-automatic video description tool for Lausanne City Archives
  - Automatic detection of shots and keyframes
  - Semi-automatic detection of scenes and stories
  - Manual description of the content
  - Browsing based on the video structure

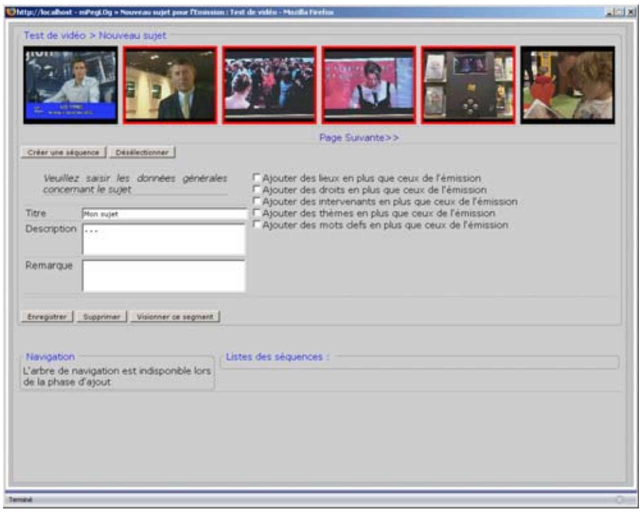
49

Video semi-automatic description: MpegLog



50

Video semi-automatic description: MpegLog (ctd.)



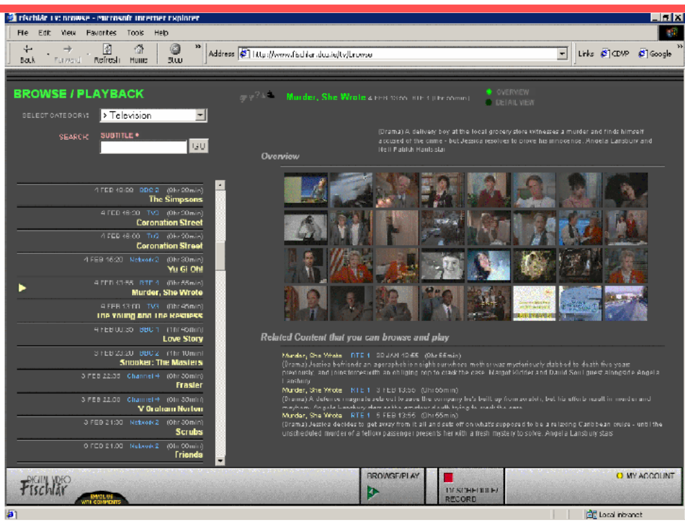
51

Video browsing

- Example: Físchlár-TV, Centre for Digital Video Processing, DCU, Ireland.
  - Supports recording, analysis, browsing, and playback of digital TV video, from 8 channels.
  - Users select programmes from a TV schedule with programme genre automatically assigned.
  - At transmission time, the systems captures video, detects shots, scenes & Keyframes and places videos in a library of content.
  - Users browse programme genres or otherwise locate programmes, and select a program for viewing;
  - Initially, users are allowed to browse Keyframes and then playback;

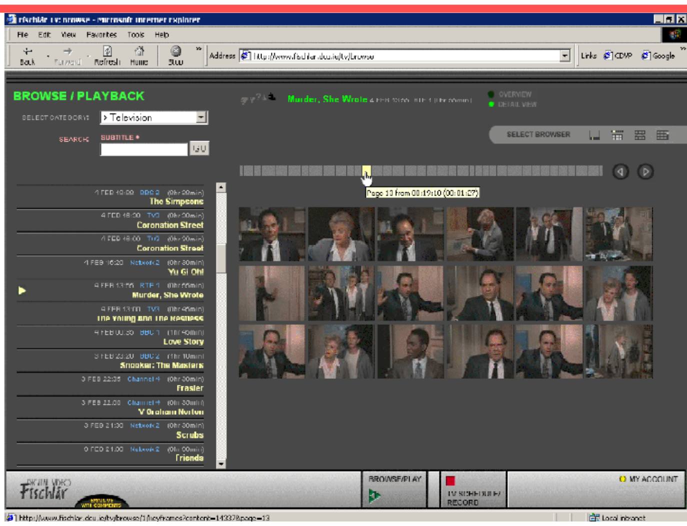
52

Video browsing: Físchlár-TV



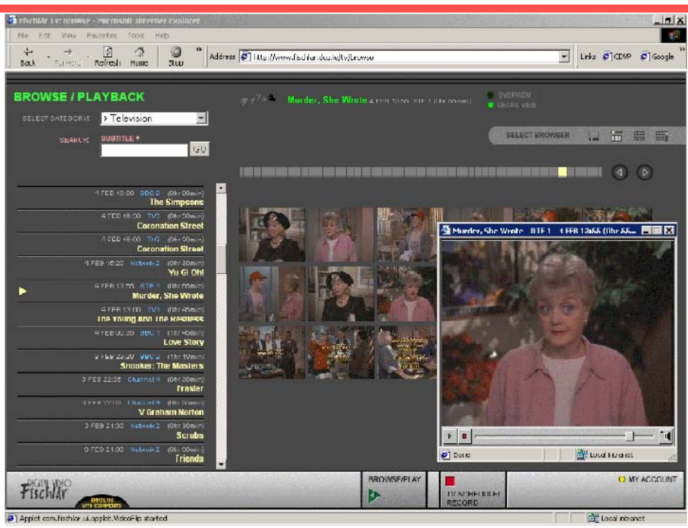
53

Video browsing: Físchlár-TV



54

Video browsing: Físchlár-TV



55

Multimedia retrieval: discussion

- Recent approaches to the problem of multimedia IR are mostly based on the extraction of text/audiovisual features
- Extraction/creation of descriptions is complex and expensive
  - Manual approaches are time consuming
  - Automatic approaches are not always possible, some are not sufficiently accurate
- Multimedia descriptions are very precious
  - Applications need to exchange them
  - Created descriptions should be conserved

=> Important need for a standard multimedia description language

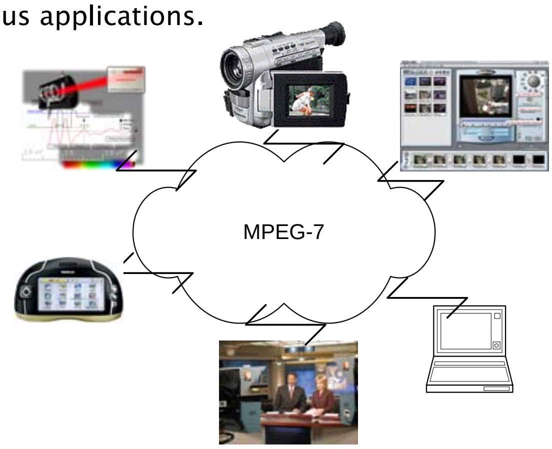
56

Standard multimedia description language : MPEG-7

- MPEG-7 is a member of MPEG standard family
  - International standard: October 2001
  - is formally defined as “Multimedia Content Description Interface”.
  - standardises the description of various types of multimedia information.
  - does not comprise the (automatic) extraction of descriptors, nor does it specify the search engine that can make use of the description.

Standard multimedia description language : MPEG-7 (ctd.)

- MPEG-7 provides a standard library for audiovisual data description in order to facilitate their exchange between various applications.



MPEG standards family

MPEG-21 Content access

MPEG-1 / 2 / 4
Content coding

MPEG-7
Content description

```
<mpeg7>
  <Descriptor>
    <MultimediaContent
xsi:type
      ...
```


Example of MPEG-7 description (I)

TV news image

Example of MPEG-7 description (II)

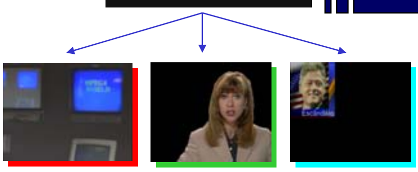
Type

```
<Mpeg7>
  <StillRegion id = “news”>

  </StillRegion>
</Mpeg7>
```


Example of MPEG-7 description (III)

Spatial decomposition



```
<Mpeg7>
  <StillRegion id = “news”>
    <SegmentDecomposition
     decompositionType = “spatial”>
      <StillRegion id =“background”>
      <StillRegion id =“speaker”>
      <StillRegion id =“topic”>
    </SegmentDecomposition>
  </StillRegion>
</Mpeg7>
```

Example of MPEG-7 description (IV)

Background features

```
<Mpeg7>
 <StillRegion id = “news”>
   <SegmentDecomposition
    decompositionType = “spatial”>
     <StillRegion id = “background”>
       <DominantColor> 110 108 140
   </DominantColor>
     <StillRegion id = “speaker”>
     <StillRegion id = “topic”>
   </SegmentDecomposition>
 </StillRegion>
</Mpeg7>
```

63

Example of MPEG-7 description (V)

More features

```
<Mpeg7>
 <StillRegion id = “news”>
   <SegmentDecomposition
    decompositionType = “spatial”>
     <StillRegion id = “background”>
     <StillRegion id = “speaker”>
       <TextAnnotation>
         <FreeTextAnnotation> Journalist Judite  Sousa
         </FreeTextAnnotation>
       </TextAnnotation>
       <SpatialMask>
         <Poly>
           <Coords> 80 288, 100 200, ..., 352 288 </Coordsl>
         </Poly>
       </SpatialMask>
     <StillRegion id = “topic”>
   </SegmentDecomposition>
 </StillRegion>
</Mpeg7>
```

64

MPEG-7 : discussion

- Recent standard, not yet widely implemented in commercial applications
- A few MPEG-7 compatible research applications:
  - ex: Caliph and Emir http://www.semanticmetadata.net/features/
- Very big standard: we need to determine useful parts for a specific application: profiles.
- Great potential for multimedia reuse, indexing and retrieval.

65

Conclusions: Promising directions for MMIR

- Multi-modal indexing/retrieval
  - Combining different features, using machine learning approaches to model the combination of features into higher-level semantics (INFORMEDIA project, CMU)
  - Query-class-dependency: classification of queries and resolving them using specific class-dependent tools (Prof. Shih-Fu Chang, DVMM lab, Columbia University, NY.)

66

Conclusions: Promising directions for MMIR

- Using *folksonomies*
  - A folksonomy is a user generated taxonomy used to categorize and retrieve web content such as Web pages, photographs and Web links, using open-ended labels called tags.
  - Folksonomies can help to provide collaborative description of media resources and therefore facilitate access to their content.
  - Examples of use: flickr.com et del.icio.us.

67

Conclusions: Promising directions for MMIR

- Multimedia data mining
  - Enriching descriptions based on correlations between concepts
    - Using large scale video description corpus
      - TRECVid: TREC Video Retrieval Evaluation (http://www-nlpir.nist.gov/projects/t01v/)
      - LSCOM: Large Scale Concept Ontology for Multimedia (http://www.ee.columbia.edu/ln/dvmm/lscom/)
    - Discovering association rules between annotated concepts.
    - Applying rules to a partially described content to derive new descriptions.
    - See LSVAM project (HEIG-VD and DCU).

68

Conclusions: Promising directions for MMIR

- *Be imaginative!*

  As the most effective solutions are not necessarily the most complex ones.

- Google currently proposes a fun way of image indexing.
  - Players on the Web are presented an image in parallel. If two players propose the same description for the image, they gain points…
  - See http://images.google.com/imagelabeler/

69

Thank you for your attention!

70