MPEG-7

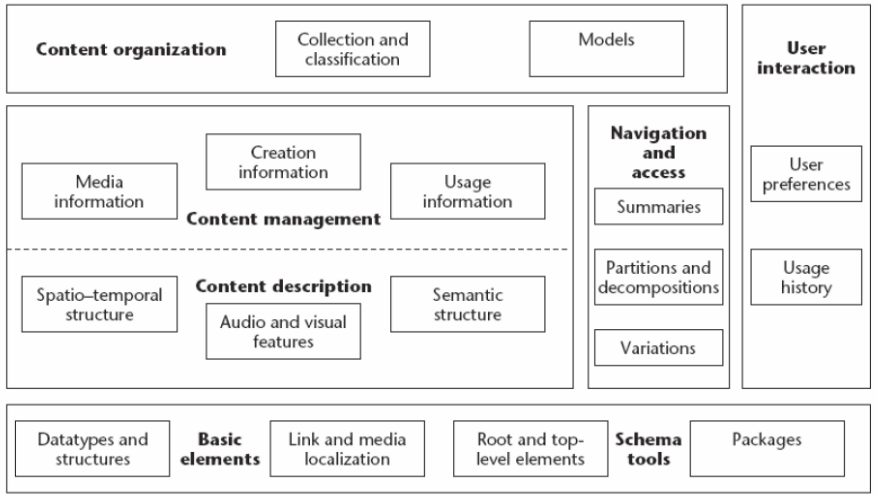
Multimedia Description Schemes (MDS)

Learning Objectives

- Describe different elements in multimedia description schemes
- Introduction to the XML-based schema
- MPEG-7 example application
- References:
  - Introduction to MPEG-7, Multimedia Content Description Interface, John Wiley & Sons, 2002.
  - http://muscle.isti.cnr.it/Standards/MPEG-7/sw-tools.xml

Multimedia Description Schemes

- Specifies high level framework for generic descriptions of all kinds of multimedia
  - Description tools
    - Metadata structures for describing and annotating multimedia content
- Levels:
  - Basic elements
  - Content management and content description



Basic Elements

- Essentials of multimedia content description
- Used repeatedly in multimedia content descriptions
- Examples:
  - Basic data types: e.g., Matrix
  - Time, duration
  - Text annotation

DS for Content Management

- Creation and Production:
  - Title, creator, creation location and date, parental guidance, …
- Content Usage:
  - Rights holder, usage rights, usage record, …
- Media description:
  - Compression, coding and storage format of multimedia content

DS for Content Management

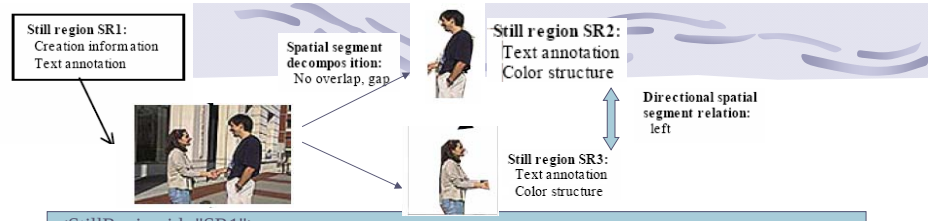
```
<CreationInformation>
  <Creation>
        <Creator>
              <Role><Name xml:lang="en">Photographer</Name></Role>
              <Agent xsi:type=”PersonType”>
                  <Name> <GivenName>Somebody</GivenName>  </Name>
              </Agent>
        </Creator>
        <CreationCoordinates>
              <Location>
                    <Name xml:lang="en">Columbia University</Name>
                    <Region>us</Region>
              </Location>
              <Date>  <TimePoint>1998-09-19</TimePoint>  </Date>
        </CreationCoordinates>
  </Creation>
</CreationInformation>
```


DS for Content Management

```
<MediaFormat>
        <Content href=" urn:mpeg:mpeg7:cs:ContentCS:2001:1">
              <Name xml:lang="en">image</Name>
        </Content>
        <FileFormat href=" urn:mpeg:mpeg7:cs:FileFormatCS:2001:1">
              <Name xml:lang="en">jpeg</Name>
        </FileFormat>
        <VisualCoding>
              <Format colorDomain="color"
                    href=" urn:mpeg:mpeg7:cs:VisualCodingFormatCS:2001:1">
                    JPEG
              </Format>
              <Frame height="480" width="704"/>
        </VisualCoding>
</MediaFormat>
```

Content Description

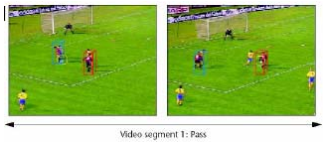
- Describe the structure and semantics of multimedia data
- Structure:
  - Segments:
    - Objects in an image, video short, audio segment
    - How: use low level descriptors and text annotation
  - Example:
    - A single image: decomposed into a number of regions
    - Each image region can be described using different descriptors




```
<StillRegion id="SR1">
  <TextAnnotation>
    <FreeTextAnnotation>  Alex shakes hands with Ana   </FreeTextAnnotation>
  </TextAnnotation>
  <SpatialDecomposition overlap="false" gap="true">
    <StillRegion id="SR2">
      <TextAnnotation> <FreeTextAnnotation> Alex </FreeTextAnnotation>
      </TextAnnotation>
      <VisualDescriptor xsi:type="ColorStructureType"> ... </VisualDescriptor>
    </StillRegion>
    <StillRegion id="SR3">
      <TextAnnotation> <FreeTextAnnotation> Ana </FreeTextAnnotation>
      </TextAnnotation>
      <Relation xsi:type="DirectionalSpatialSegmentRelationType“ name="left“
                target="#SR2"/>
      <VisualDescriptor xsi:type="ColorStructureType"> ... </VisualDescriptor>
    </StillRegion>
  </SpatialDecomposition>
</StillRegion>
```

Content Description

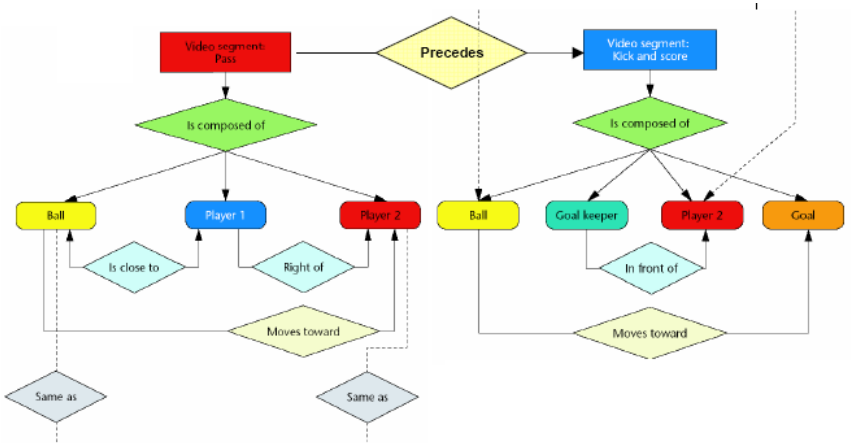
- Example: a video clip
  - Decomposed onto two segments:
    - Video segment 1: Pass



    - Video segment 2: Kick and score

- Each video segment can be decomposed into various moving regions

- Each moving region
  - Descriptors?
  - Structural relation

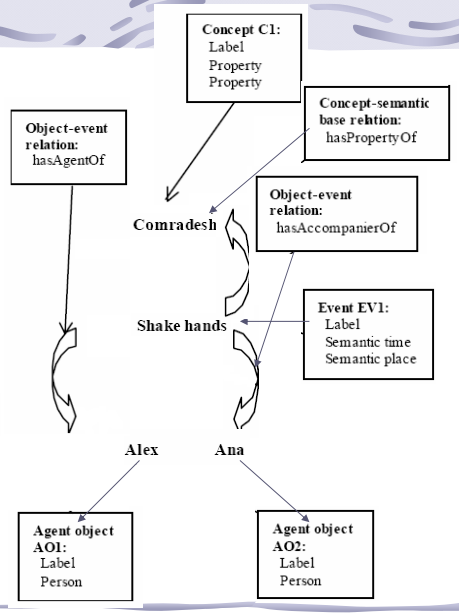


Content Description

- Describe the structure and semantics of multimedia data
- Semantics
  - Describing semantic entities in the narrative world
    - People, actions, relation between people and actions, …
    - Events: perceivable event that takes place
    - Semantic relation: general relations between entities
      - "hasAgentOf" – initiates the action of an event
      - "HasAccompanierOf" – object that is a join agent in an event

Example

- Semantics:
  - Location?
  - Time?



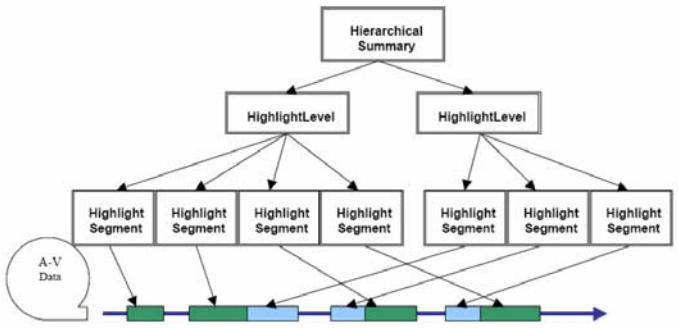

```
<Semantic>
    <Label><Name>Alex shakes hands with Ana </Name></Label>
    <SemanticBase xsi:type="EventType" id="EV1">
            <Label><Name>Shake hands</Name></Label>
            <Relation xsi:type="ObjectEventRelationType“  name="hasAgentOf“
                    target="#AO1"/>
            <Relation xsi:type="ObjectEventRelationType“ name="hasAccompanierOf“
                    target="#AO2"/>
            <Relation xsi:type="ConceptSemanticBaseRelationType“ name="hasPropertyOf"
                    target="#C1"/>
            <SemanticPlace> <Label><Name>Columbia University</Name></Label>
            </SemanticPlace>
            <SemanticTime> <Label><Name>September 9,1998</Name></Label>
            </SemanticTime>
    </SemanticBase>
```

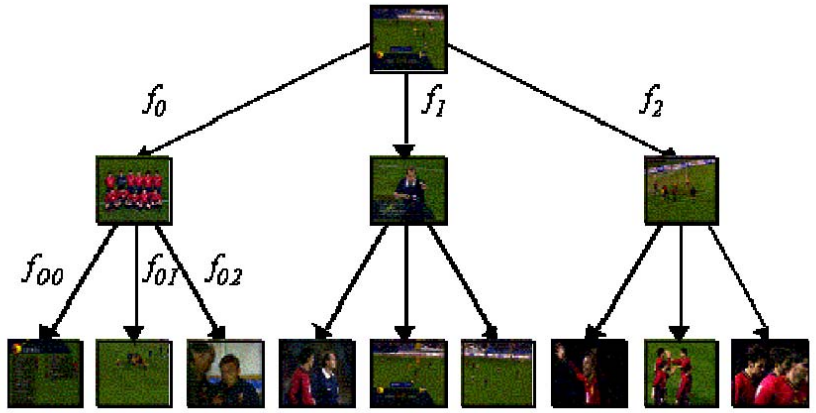
```
<SemanticBase xsi:type="AgentObjectType" id="AO1">
        <Label><Name>Alex</Name></Label>
        <Agent xsi:type="PersonType">
                <Name><GivenName>Alex</GivenName></Name> </Agent>
</SemanticBase>
<SemanticBase xsi:type="AgentObjectType" id="AO2">
        <Label><Name>Ana</Name></Label>
        <Agent xsi:type="PersonType">
                <Name><GivenName>Ana</GivenName></Name> </Agent>
</SemanticBase>
<SemanticBase xsi:type="ConceptType" id="C1">
        <Label><Name>Comradeship</Name></Label>
                <Property>Associate</Property>
                <Property>Friend</Property>
</SemanticBase>
</Semantic>
```


Navigation and Access

- Description schemes for enabling browsing and retrieval
  - Summarization tools:
    - Summarize a long video clip to highlight important segments
    - Allows fast browsing of content
    - Example
      - Highlights of a soccer game (just shots at goal!)

- Enables the summary of audio-video content to be captured in XML format





Navigation and Access

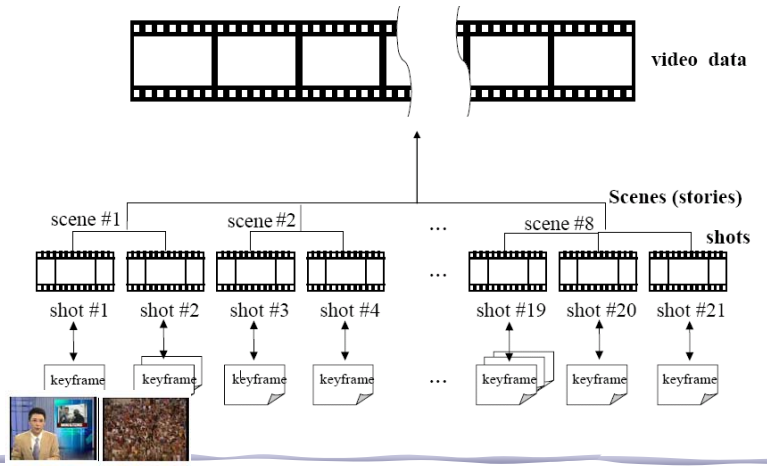
- Description schemes for enabling browsing and retrieval
  - View tools:
    - Different partitions and decompositions of image, video and audio
  - Variations:
    - Describe different variations of content available
    - Example
      - Low resolution version
      - Video only version
      - Audio only version

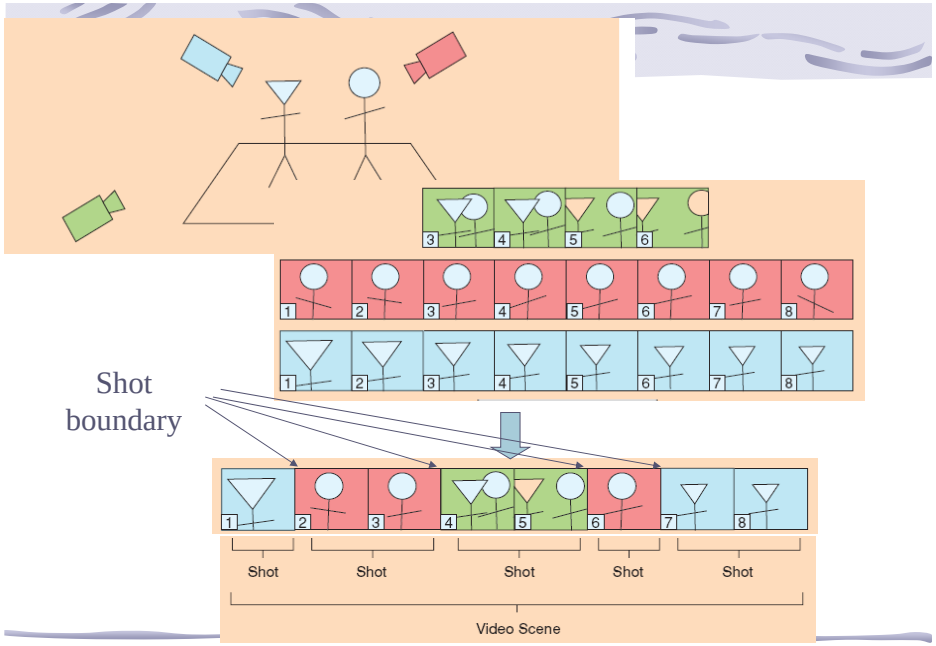
Video Structures and Summaries

Structure of a video

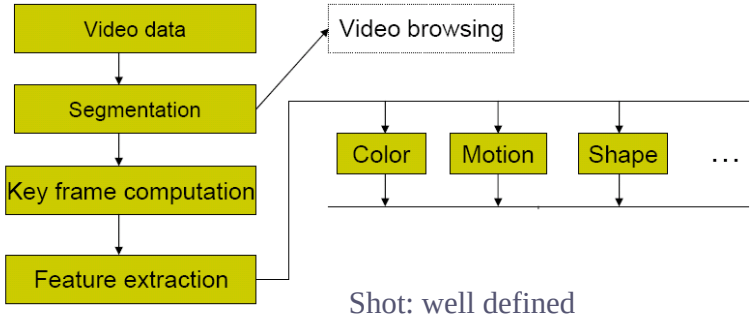
- Frame
  - The set of all picture elements that represent one image
- Shot
  - A series of interrelated consecutive frames taken continuously by a single camera and representing a continuous action in time and space
- Scene
  - A sequence of shots with a common object of interest, a common location or a common thematic concept
- http://classes.yale.edu/film-analysis/htmfiles/editing.htm

Structure of a video





Structure of a video



Shot: well defined

Key frame: representative frame inside a shot

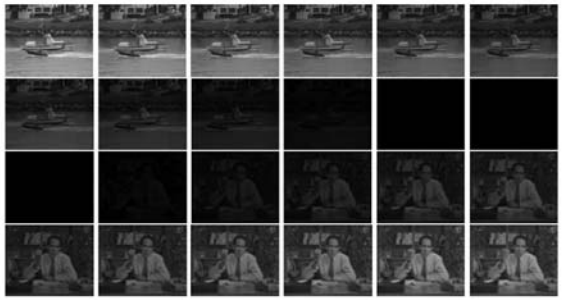
How to decompose a Video?

- Shot segmentation
  - Detect shot boundaries
  - Examples:
    - Cut
      - Abrupt transition between two shots

    - Fade
      - Fade out and fade in

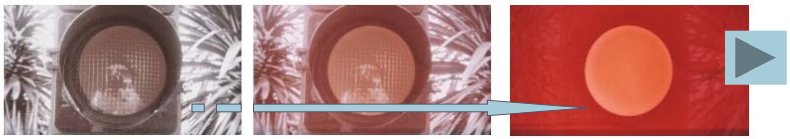
How to decompose a Video?

- Major types:
  - Fade
    - E.g., a fade out to black: the end of a scene
    - Slow decrease/increase in brightness



Different types of editing

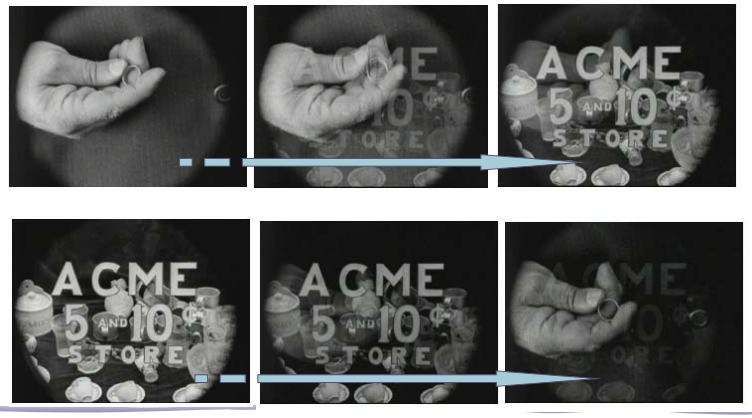
- Video structure: Scene + shots
- Abrupt cuts: High motion
- Smooth transition between 2 shots



- Abrupt + dissolve

Different types of editing

- Abrupt + dissolve



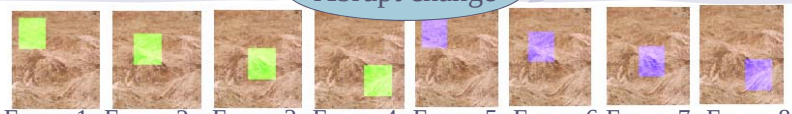
Shot Boundary Detection

- Frame by Frame Difference
  - Small difference between similar frames
  - Big difference in the shot boundary

Frame t | frame t spatially shifted by 1 | difference image

- Motion compensated difference

Abrupt change



Frame 1 Frame 2 Frame 3 Frame 4 Frame 5 Frame 6 Frame 7 Frame 8

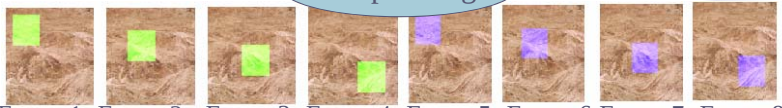
8-bin histogram

	R								G								B							
frame 1	0	0	0	0	4	30	57	9	0	0	1	9	39	36	5	10	0	1	11	43	35	9	1	0
frame 2	0	0	0	0	4	30	57	9	0	0	1	10	39	35	5	10	0	1	11	43	35	9	1	0
frame 3	0	0	0	0	4	30	57	9	0	0	1	11	38	35	5	10	0	1	11	43	35	9	1	0
frame 4	0	0	0	0	4	30	57	9	0	0	1	10	38	36	5	10	0	1	11	43	35	9	1	0
frame 5	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	1	10	37	32	9	1	10
frame 6	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	0	11	36	33	9	1	10
frame 7	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	0	10	37	33	9	1	10
frame 8	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	0	10	38	32	9	1	10

Absolute difference

	R								G								B							
frame2 and frame1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
frame3 and frame2	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
frame4 and frame3	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
frame5 and frame4	0	0	0	0	0	0	0	0	0	0	0	1	4	4	1	10	0	0	1	6	3	0	0	10
frame6 and frame5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
frame7 and frame6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
frame8 and frame7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0

Abrupt change



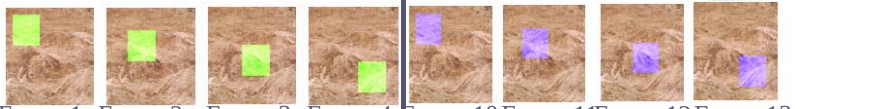
Frame 1 Frame 2 Frame 3 Frame 4 Frame 5 Frame 6 Frame 7 Frame 8

	R								G								B							
frame2 and frame1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
frame3 and frame2	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
frame4 and frame3	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
frame5 and frame4	0	0	0	0	0	0	0	0	0	0	0	1	4	4	1	10	0	0	1	6	3	0	0	10
frame6 and frame5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
frame7 and frame6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
frame8 and frame7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0

Total Absolute difference

	R	G	B	
frame2 and frame1	0	2	0	2
frame3 and frame2	0	2	0	2
frame4 and frame3	0	2	0	2
frame5 and frame4	0	20	20	40
frame6 and frame5	0	0	4	4
frame7 and frame6	0	0	2	2
frame8 and frame7	0	0	2	2

Smooth change



Frame 1 Frame 2 Frame 3 Frame 4 Frame 10 Frame 11 Frame 12 Frame 13

Dissolve

Frame 5 Frame 6 Frame 7 Frame 8 Frame 9

Smooth change

8-bin histogram

	gradual shot	R								G								B							
Shot 1	frame 1	0	0	0	0	4	30	57	9	0	0	1	9	39	36	5	10	0	1	11	43	35	9	1	0
	frame 2	0	0	0	0	4	30	57	9	0	0	1	10	39	35	5	10	0	1	11	43	35	9	1	0
	frame 3	0	0	0	0	4	30	57	9	0	0	1	11	38	35	5	10	0	1	11	43	35	9	1	0
	frame 4	0	0	0	0	4	30	57	9	0	0	1	10	38	36	5	10	0	1	11	43	35	9	1	0
dissolve	frame 5	0	0	0	0	4	30	57	9	0	0	1	10	38	36	15	0	0	1	11	40	38	7	1	2
	frame 6	0	0	0	0	4	30	57	9	0	0	1	10	38	46	5	0	0	1	11	39	37	6	1	5
	frame 7	0	0	0	0	4	30	57	9	0	0	1	11	42	47	4	0	0	1	11	38	37	6	1	6
	frame 8	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	1	10	37	35	8	1	8
	frame 9	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	1	10	37	33	8	1	10
shot 2	frame 10	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	1	10	37	32	9	1	10
	frame 11	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	0	11	36	33	9	1	10
	frame 12	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	0	10	37	33	9	1	10
	frame 13	0	0	0	0	4	30	57	9	0	0	1	11	42	40	6	0	0	0	10	38	32	9	1	10

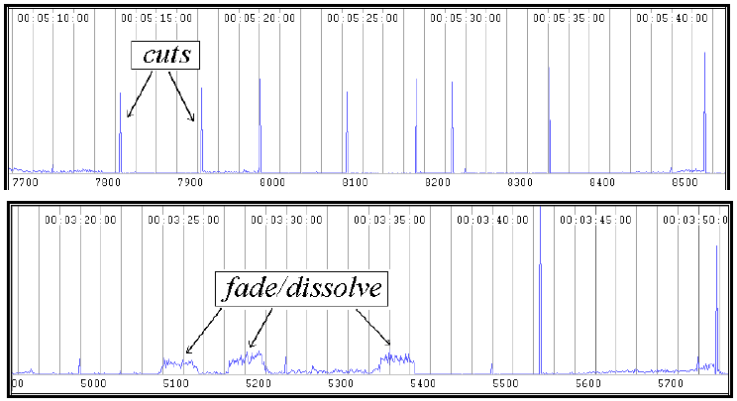
Total Absolute difference

		R	G	B	
shot 1	frame2 and frame1	0	2	0	2
	frame3 and frame2	0	2	0	2
	frame4 and frame3	0	2	0	2
dissolve	frame5 and frame4	0	20	10	30
	frame6 and frame5	0	20	6	26
	frame7 and frame6	0	7	2	9
	frame8 and frame7	0	9	8	17
	frame9 and frame8	0	0	4	4
	frame10 and frame9	0	0	2	2
shot 2	frame11 and frame10	0	0	4	4
	frame12 and frame11	0	0	2	2
	frame13 and frame12	0	0	2	2

1 2 3 4 5 6 7 8 9 10 11 12 13

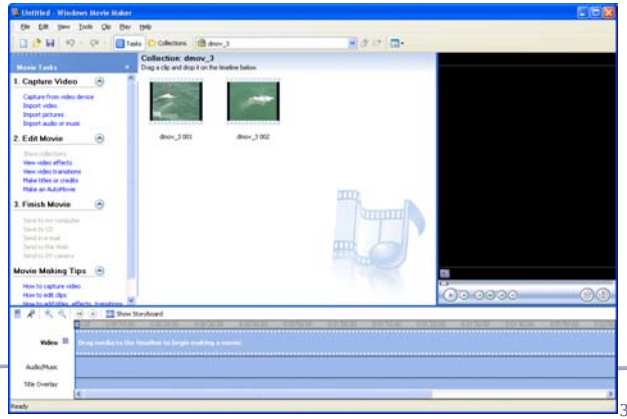
Shot Boundary Detection

- Histogram comparison



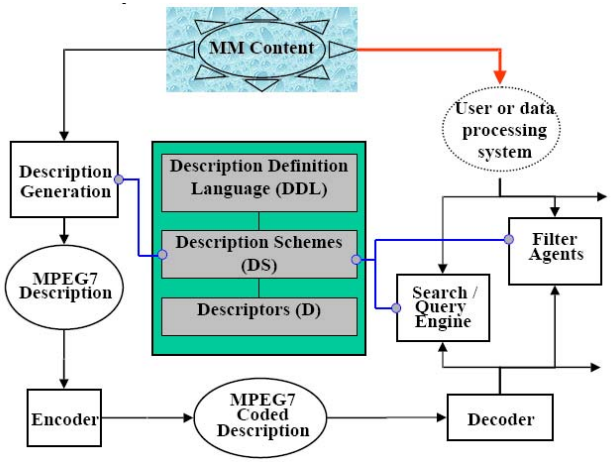
Shot Boundary Examples

- Windows Movie Maker

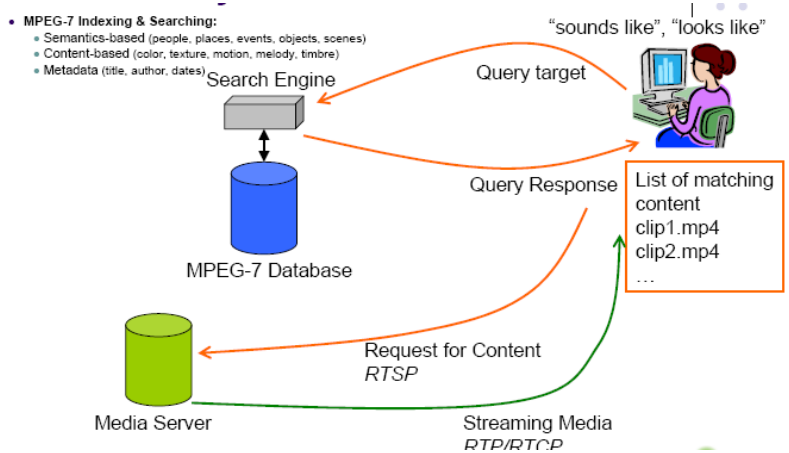


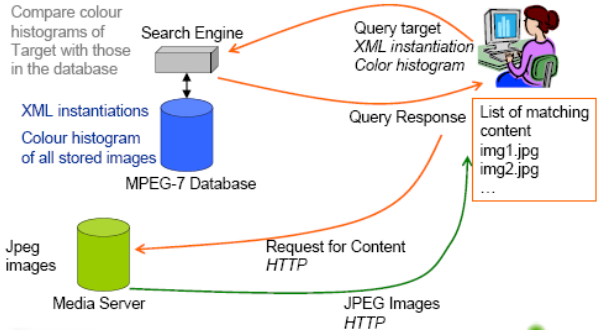
Examples

Abstract Representation



Client-server architecture






```
<VisualDescriptor xsi:type="ScalableColorType" numOfCoeff="32">
  <Coeff> 62 17 -127 47 -8 13 22 30 -31 -33 3 13
          -25 -11 13 20 2 -13 -1 3 -11 -10 1 6 2
          -1 0 0 -9 5 1 -4
  </Coeff>
</VisualDescriptor>
```


MPEG-7 Resources

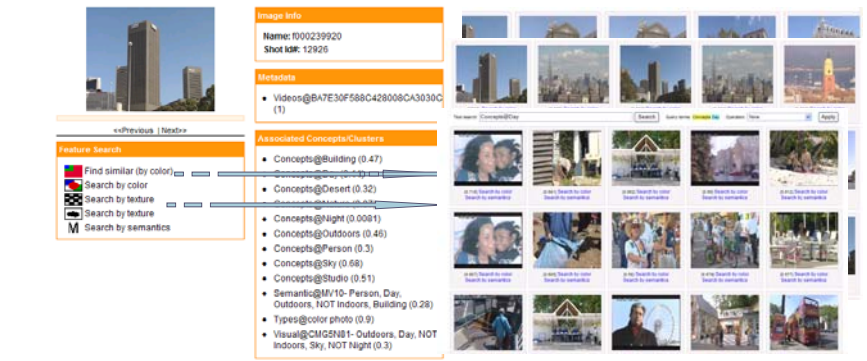
- http://muscle.isti.cnr.it/Standards/MPEG-7/sw-tools.xml
  - MPEG-7 Visual Annotation / Video Annotation and Summaries
  - MPEG-7 Library
    - A set of C++ classes, implementing the MPEG-7 standard
    - Useful to application developers
  - MARVEL: MPEG-7 Multimedia Search Engine
  - MPEG-7 Spoken Content Demonstrator
  - MPEG-7-based Audio annotation for the archival of digital video

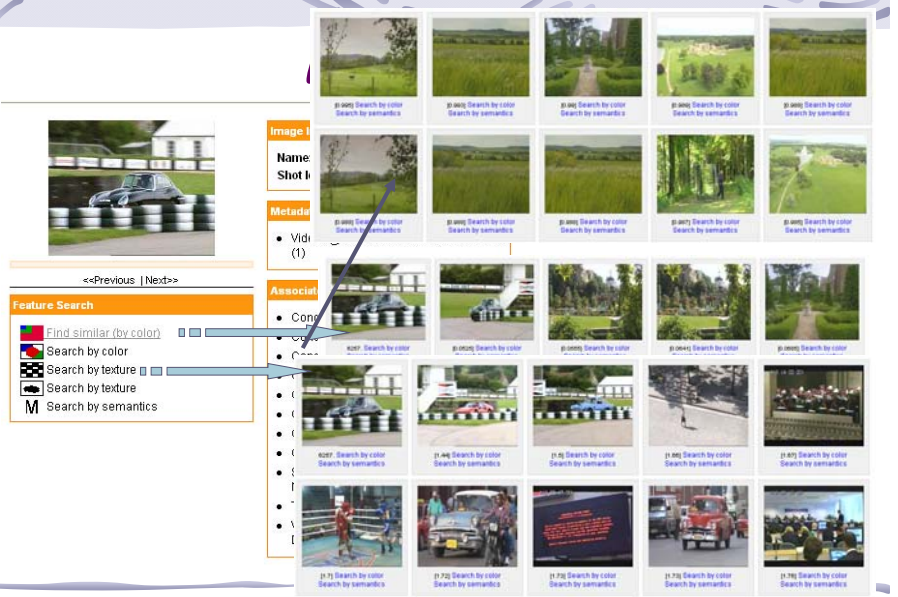
MARVEL

http://mp7.watson.ibm.com/

Search by inputting similar images

MARVEL

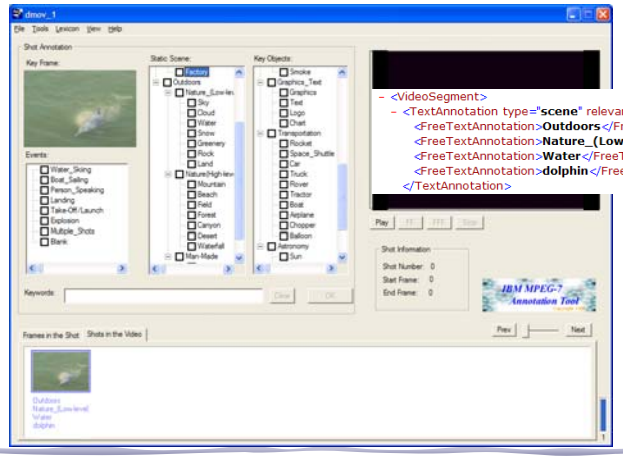




MPEG-7 Visual Annotation

- Creates MPEG-7 documents for video streams
- Describes structure and gives keywords from a *controlled* vocabulary
  - Each shot in a video sequence can be annotated with static scene descriptions, key object descriptions, …
  - Stored as MPEG-7 descriptions in an output XML file

MPEG-7 Visual Annotation

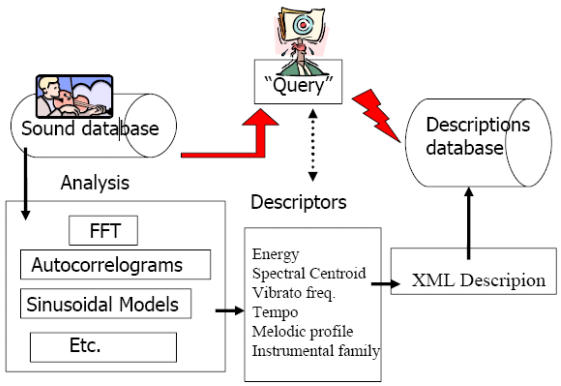


Audio Features

Content Description

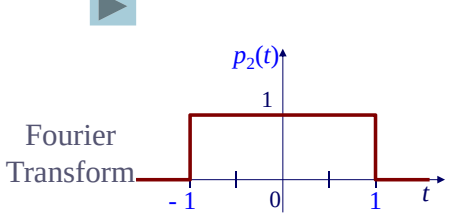
- Addresses four classes of audio
  - Pure music, pure speech, pure sound effects, arbitrary soundtracks
- Music/audio descriptors
  - Low level descriptors:
    - Fundamental frequency, amplitude, spectral deviation, ….
  - High level descriptors:
    - Instrument type, singer type, musical type, sound mix characteristics, …

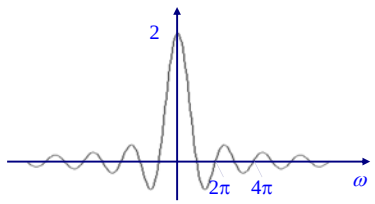
Example

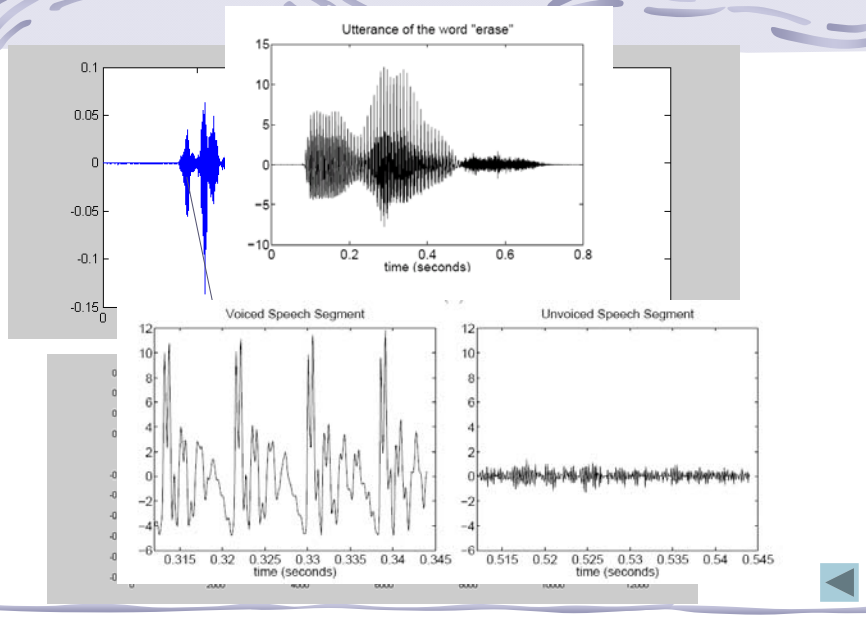


- Sound: time-varying signal: composed of
  - Low frequencies: vowels
  - High frequencies: consonants

Formant Frequencies for the Vowels				
Typewritten Symbol for the Vowel	Typical Word	F1 (Hz)	F2 (Hz)	F3 (Hz)
IY	(beet)	270	2290	3010
I	(bit)	390	1990	2550
E	(bet)	530	1840	2480
AE	(bat)	660	1720	2410
UH	(but)	520	1190	2390
A	(hot)	730	1090	2440
OW	(bought)	570	840	2410
U	(foot)	440	1020	2240
OO	(boot)	300	870	2240
ER	(bird)	490	1350	1690

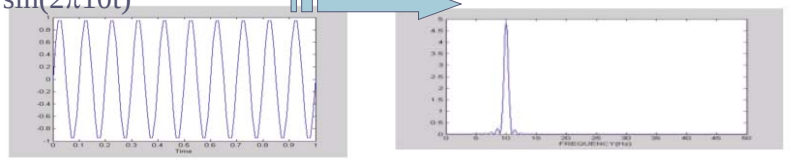




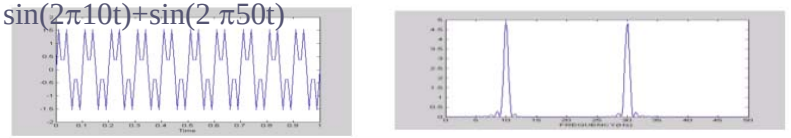


Fourier transform

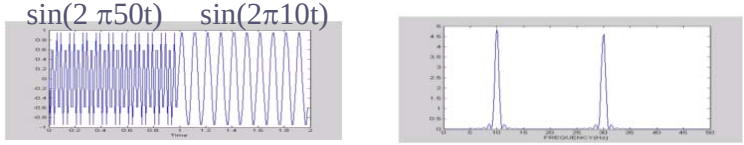
$\sin(2\pi 10t)$



$\sin(2\pi 10t)+\sin(2\pi 50t)$

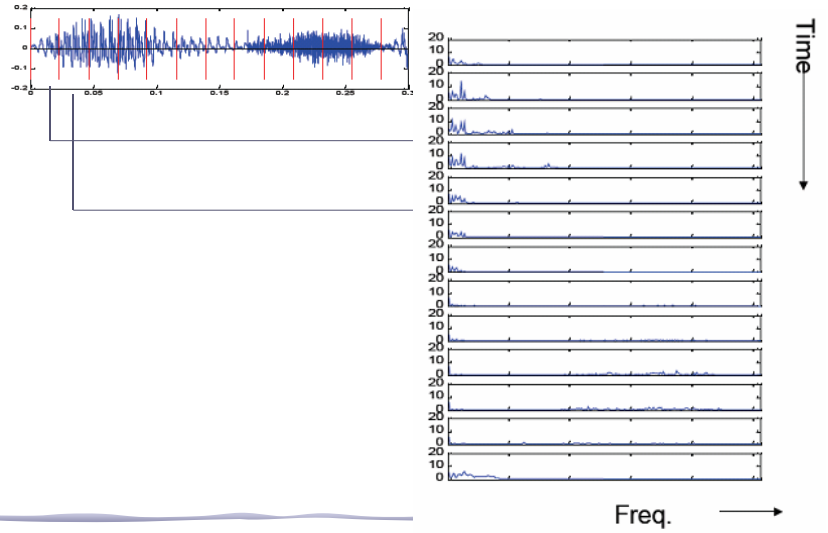


$\sin(2\pi 50t)$ $\sin(2\pi 10t)$

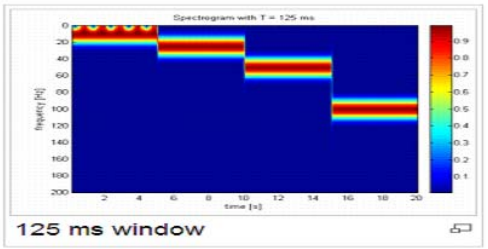


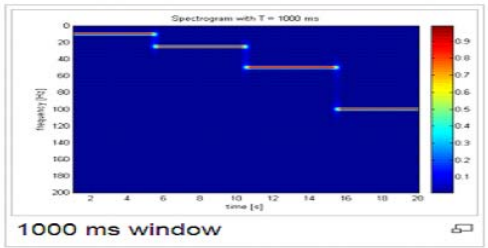
No time information in frequency domain

Windowing

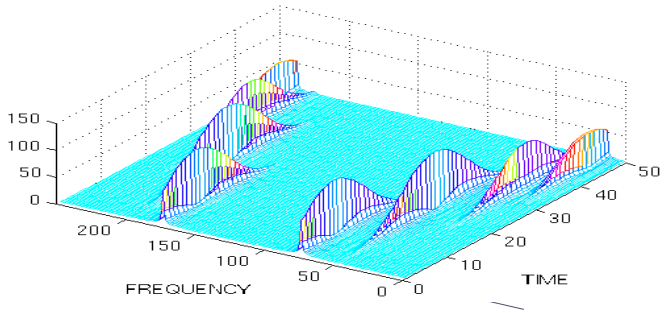


$$x(t) = \begin{cases} \cos(2\pi 10t); & 0 \le t < 5s \\ \cos(2\pi 25t); & 5 \le t < 10s \\ \cos(2\pi 50t); & 10 \le t < 15s \\ \cos(2\pi 100t); & 15 \le t < 20s \end{cases}$$



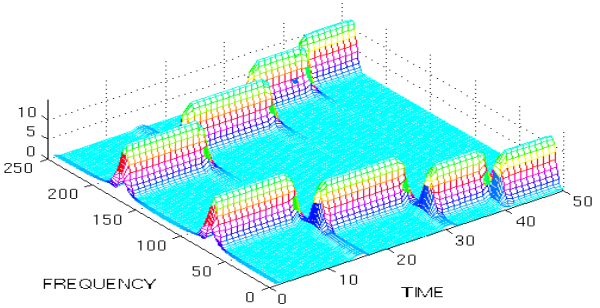


Long window: good frequency resolution



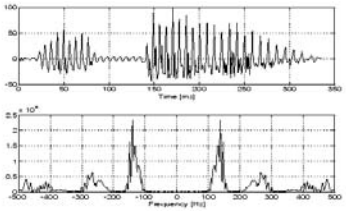
Four peaks are well separated in frequency, but not in time.

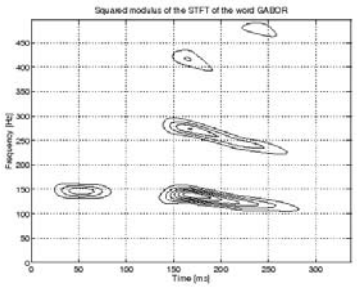
Short window: good time resolution

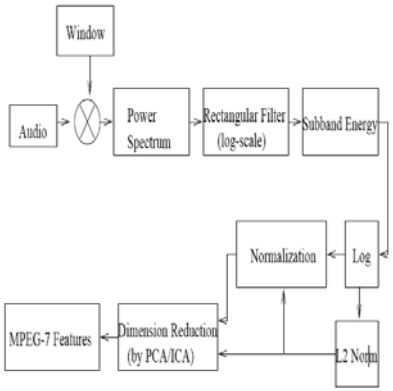


Four peaks are well separated in time, but each covers a range of frequencies

Time-frequency Analysis: Word: Gabor





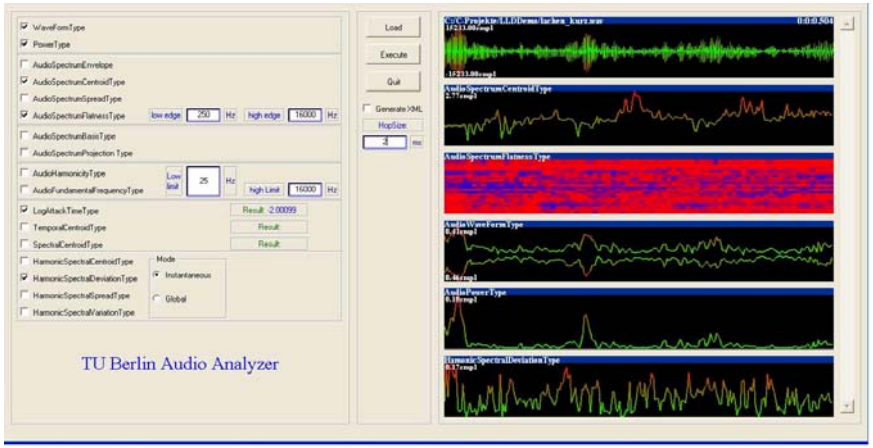


MPEG7 Example Audio Descriptors

- Silence Description
  - "no significant sound"
  - Help to further segment the audio stream
- Basic
  - Audio waveform description
    - Describes audio waveform envelope (minimum and maximum)
  - Audio power description
    - Power spectrum
- Basic Spectral: all from time-frequency analysis of an audio signal
  - Audio spectrum envelope description
    - Logarithmic frequency spectrum: describes short term power spectrum of an audio signal
  - Audio spectrum centroid description
    - Describes the shape of the power spectrum
    - Indicates whether the spectral content is dominated by high or low frequencies
  - Audio spectrum flatness description
    - Describes the flatness properties of the spectrum

MPEG7 Audio Descriptors

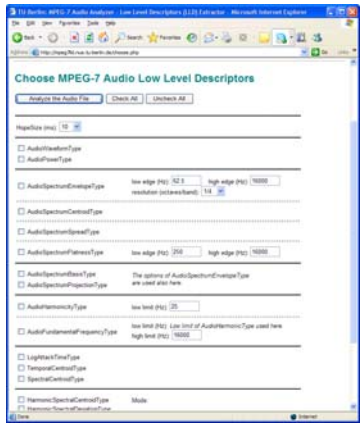
- Spectral Basic
  - Audio spectrum basis description
    - Mainly used for sound classification
- Signal Parameters
  - Audio fundamental frequency description
    - Describes the fundamental frequency
    - Noise segment: no fundamental frequency
  - Audio harmonicity description
    - Musical tones/voided speech: harmonic spectrum
    - Noise, unvoiced speech: non-harmonic spectrum



MPEG-7 Spoken Content Demonstrator

- http://mpeg7lld.nue.tu-berlin.de/
- Creating MPEG-7 documents with SpokenContent description from an input speech signal (WAV, MP3)

MPEG-7 Spoken Content Demonstrator

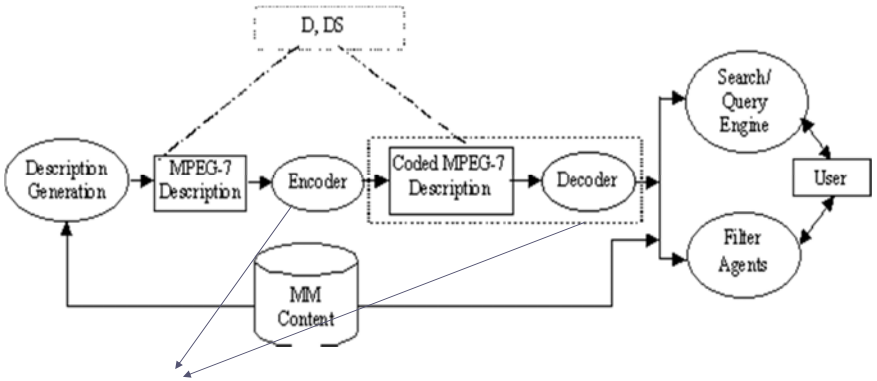


XML file

What you have learnt?

- Understand the purposes of MPEG-7.
- Understand differences between
  - low level and
  - high level descriptors
- Understand different applications areas

Abstract Representation



Optional

Signal Processing in Multimedia Systems

- Compression of multimedia signals
  - Lossless and lossy approaches
  - Consider signals characteristics
    - Discard unimportant info
    - Take advantage of human visual/hearing characteristics
- Accessing, searching, browsing multimedia signals
  - Based on low level features extractable from the multimedia signals
- Multimedia applications: standard based