Mpeg 7

Content description
Mpeg 7
ch. 9.5
Content, content, and more content …
How to get what is needed?

- Increasing availability of multimedia information
- Difficult to find, select, filter, manage AV content
- More and more situations where it is necessary to have ‘information about the content’
Goals for Multimedia Content Retrieval

- Because the value of content depends on how easy it is to find, filter, manage, and use it.
- Want to make multimedia content searchable like text information
- Need content description method beyond simple text annotation
Architecture for Multimedia Retrieval

- Feature extraction
  - Manual / automatic
- AV Description
- Storage
- Transmission
  - Decoding (for transmission)
  - Encoding (for transmission)

- Search / query
- Browse
- Filter

Push / Pull

Conf. points

Human or machine
MPEG Standards

- MPEG-1: For the storage and retrieval of moving pictures and audio on storage media.
- MPEG-2: For digital television, it’s the timely response for the satellite broadcasting and cable television industries in their transition from analog to digital formats.
- MPEG-4: Codes content as objects and enables those objects to be manipulated individually or collectively on an audiovisual scene.
- MPEG-1, -2, and -4 make content available.
- MPEG-7 lets you to find the content you need.
<table>
<thead>
<tr>
<th>Layer</th>
<th>OSI-Definition</th>
<th>Expl.: Written Message</th>
<th>MPEG Layers</th>
<th>MPEG Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Application</td>
<td>read and understand</td>
<td>Content Layer</td>
<td>MPEG 7</td>
</tr>
<tr>
<td>6</td>
<td>Presentation</td>
<td>Arrangement of writing, Structure</td>
<td>Composition Layer</td>
<td>MPEG 4</td>
</tr>
<tr>
<td>5</td>
<td>Session</td>
<td>Alphabet, Language (English, Italiano)</td>
<td>Compression Layer</td>
<td>MPEG 1/2</td>
</tr>
<tr>
<td>4</td>
<td>Transport</td>
<td>Paper, Ink</td>
<td>System Layer</td>
<td>MPEG 1/2</td>
</tr>
<tr>
<td>3</td>
<td>Network</td>
<td>Postbox</td>
<td></td>
<td>DVB</td>
</tr>
<tr>
<td>2</td>
<td>Data Link</td>
<td>Address on Letter</td>
<td></td>
<td>DVB</td>
</tr>
<tr>
<td>1</td>
<td>Physical</td>
<td>Transport through carrier</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MPEG-7: What Is It?

**THE MPEG 7 STANDARD**

- IS NOT a STANDARD for FEATURE EXTRACTION/ MATCHING
- IS NOT a COMPRESSION Standard similar to MPEG-1/2/4 or their Extension

Content Description of Various Audio Visual Information

**Types of Audio Visual Information**
- Audio, Speech
- Moving video, still pictures, graphics
- Information on how objects are combined in scenes
Why do we need MPEG-7?

Need

- Fast & Accurate Access
- Personalized Content Production and Consumption
- Content Management
- Automation

Support for Advanced Query

- Visual
- Audio
- Sketch
Standardize object-based description tools for various types of audiovisual information, allowing fast and efficient content searching, filtering and identification, and addressing a large range of applications.

New objective for MPEG:
- MPEG-1, -2 and -4 represent the content itself (‘the bits’)
- MPEG-7 should represent information about the content (‘the bits about the bits’)
Most types of audiovisual information are considered:
- Audio, speech;
- Moving video, still pictures, graphics, 3D models;
- Information on how objects are combined in scenes.

Definitions independent of the data support
Descriptions can be associated with individual objects (could be MPEG-4 objects)
Existing solutions for textual content description will be considered
Content and Description Relation

Description may be separated from the content

Description may be multiplexed with the content
Types of Applications

- **PULL Applications:** “Search and Browsing”
  - Example: Search engines for internet and databases
  - Advantage: Many search engines work on standardized descriptions

- **PUSH Applications:** “Filtering”
  - Example: Broadcast of video, Interactive TV
  - Advantage: Intelligent agents filter standardized descriptions
Example of application areas

- Storage and retrieval of audiovisual databases (image, film, radio archives)
- Broadcast media selection (radio, TV programs)
- Surveillance (traffic control, surface transportation, production chains)
- E-commerce and Tele-shopping (searching for clothes / patterns)
- Remote sensing (cartography, ecology, natural resources management)
- Entertainment (searching for a game, for a karaoke)
- Cultural services (museums, art galleries)
- Journalism (searching for events, persons)
- Personalized news service on Internet (push media filtering)
- Intelligent multimedia presentations
- Educational applications
- Bio-medical applications
Example of queries

- **Text (keywords):**
  - Find AV material with subject corresponding to some keywords

- **Semantic description:**
  - Find AV material corresponding to a specified semantic

- **Image as an example:**
  - Find an image with similar characteristics (global or local)

- **A few notes of music:**
  - Find corresponding musical pieces or movies

- **Low level features (example: motion):**
  - Find video with specific object motion trajectories
Type of Description

- Related to information that cannot be deduced from the content:
  - e.g. recording date & conditions, author, copyright data, conditions for access, context of recording, coding format, links to other relevant material

- Related to information that is present in the content:
  - LOW LEVEL FEATURES allowing automatic extraction but without semantic value
  - HIGH LEVEL (SEMANTIC) FEATURES related to the human interpretation of the content

A ‘good’ description strongly depends on the application!
Type of descriptions

- Low level description (features, etc.)
  - Generic and flexible
  - Intelligent / efficient search engine

- High level description (structures, concepts, etc.)
  - Efficient and powerful
  - Lack of flexibility

Low-level and high-level querying may co-exist.
Low-level Description

- Information in the creation and production processes
  - *director, title, short feature movie*
- Information related to the usage of the content
  - *copyright pointers, usage history, broadcast schedule*
- Information on the storage features of the content
  - *storage format, encoding*
- Information about low-level features in the content
  - *colours, textures, sound timbres, melody*
High-level Description

- **Structural description**
  - video segments, frames, still and moving regions, audio segments
  - Segment DS (representing the spatial, temporal or spatio-temporal structure)

- **Conceptual (semantic) description**
  - objects, events, and notions
  - links of the two descriptions
Defining MPEG-7

- Comprehensive set of AV descriptions based on:
  - Catalogue (e.g. title, creator, rights)
  - Semantic (eg. The who, what, when, information about objects and events)
  - Structural (eg. The color histogram of an image, timbre of a recorded instrument)

- Interoperability
  - MPEG-7 uses XML Schema as language for content description
  - MPEG-7 will be interoperable with other leading standard (e.g. SMPTE)
  - Metadata Dictionary, Dublin Core, EBU P/Meta
Main Elements of MPEG-7

- **Descriptors (D)**
  - syntax and semantics of each feature representation

- **Description Schemes (DS)**
  - structure and semantics of the relationships between components

- **Description Definition Language (DDL)**
  - creation of new DS’s
  - modification/extension of existing DS’s

- **Systems tools**
  - synchronization, transmission...

---

In the standard

Not in the standard
Defined with DDL
MPEG-7 Working Areas

Descriptors: (Syntax & semantic of feature representation)

Description Schemes

Description Definition Language

Definition

Tags

Instantiation

Encoding & Delivery

⇒ extension

<scene id=1>
<time>....
<camera>..
<annotation
</scene>
MPEG-7 Framework

Description Generation

MPEG7 Description

Encoder

MPEG7 Description Definition Language (DDL)

MPEG7 Description Schemes (DS) & Descriptors (D)

Search / Query Engine

Decoder

Filter Agents

User

MM Content

Filter Agents

Search / Query Engine

Decoder

MPEG7 Coded Description

MPEG7 Description

Description Generation

Encoder
**Descriptor (D)**: A Descriptor is a representation of a Feature. A Descriptor defines the syntax and the semantics of the Feature representation.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>Histogram of Y,U,V components</td>
</tr>
<tr>
<td>Shape</td>
<td>ART moments</td>
</tr>
<tr>
<td>Motion</td>
<td>Motion field, coefficients of a model</td>
</tr>
<tr>
<td>Audio frequency</td>
<td>Average frequency components</td>
</tr>
<tr>
<td>Title</td>
<td>Text</td>
</tr>
<tr>
<td>Annotation</td>
<td>Text</td>
</tr>
<tr>
<td>Genre</td>
<td>Text, index in as thesaurus</td>
</tr>
</tbody>
</table>
Visual Features ...

Colour  Shape  Texture
Description Scheme (DS): A Description Scheme specifies the structure and semantics of the relationships between its components, which may be both Descriptors and Description Schemes.
**Description:**
- A Description consists of a DS (structure) and the set of Descriptor Values (instantiations) that describe the Data.

**Coded Description:**
- A Coded Description is a Description that has been encoded in binary form to fulfill relevant requirements such as compression efficiency, error resilience, random access, etc.

(textual and binary forms are normative)
Description Definition Language (DDL):

The DDL is the language to specify description schemes and possibly descriptors.

In the standard:
- DS
- D

Not in the standard:
- D

Defined with DDL:
- DS
- D
Description Definition Language is ...

- XML Schema (W3C) + Mpeg-7 extensions
- Reusable Mpeg-7 Schema
  - Importing: to import type declarations of Mpeg-7 schemes
  - Redefinition: To modify existing D’s or DS’s
  - Restriction: To restrict certain aspects of existing D’s or DS’s
  - Extension: To extend existing D’s or DS’s
Low level Audio Visual descriptors

Video segments
- Color
- Camera motion
- Motion activity
- Mosaic

Moving regions
- Color
- Motion trajectory
- Parametric motion
- Spatio-temporal shape

Still regions
- Color
- Shape
- Position
- Texture

Audio segments
- Spoken content
- Spectral characterization
- Music: timbre, melody
MPEG-7: Visual

- Color
  - quantization, dominant, scalable, color-structure, layout, GoF/GoP color
- Texture
- Shape
  - region-based, contour-based, 3D
- Motion
  - camera motion, motion trajectory, parametric motion, motion activity
- Localization
  - spatio temporal
- Others
  - face recognition
MPEG-7: Basic Visual Structures

- Grid Layout
- 2D-3D Multiple View
- Time Series
- Spatial 2D Coordinates
- Temporal Interpolation
Mpeg-7 Audio

- **Low-level / generic tools**
  - Scale Tree (20 D’s): temporal envelope, spectral envelope, harmonicity…
  - Silence segment (1 D): levels of silence

- **Application-specific tools**
  - Sound effects (5 D’s)
  - Musical instrument timbre (3 D’s)
  - Spoken content (12 D’s)
  - Melody contour (4 D’s), Melody (5 D’s)
Example

```xml
<SoundEffectModel id="sfx1.1" SoundEffectCategoryRef="Bark">
  <ProbabilityModel xsi:type="ContinuousMarkovModelType" numberStates="7">
    0.04  0.34  0.12  0.04  0.34  0.12  0.00 </Initial>
    0.91  0.02  0.00  0.00  0.05  0.01  0.01
    0.01  0.99  0.00  0.00  0.00  0.00  0.00
  </Transitions>
</ProbabilityModel>

<AudioSpectrumBasis loEdge="62.5" hiEdge="8000" resolution="1/4 octave">
  <Matrix dim="31 5">
    0.26 -0.05  0.01 -0.70  0.44
    0.34  0.09  0.21 -0.42 -0.05
  </Matrix>
</AudioSpectrumBasis>
</SoundEffectModel>
```

- **SoundEffectCategoryRef**: a category label defined in Controlled Terms
- **ProbabilityModel**: statistical model used for content classification
- **AudioSpectrumBasis**: a projection matrix to reduce the dimensionality of a frequency spectra
Applications domains

- Architecture, real estate, and interior design (e.g., searching for ideas)
- Broadcast media selection (e.g., radio channel, TV channel)
- Cultural services (history museums, art galleries, etc.)
- Digital libraries (e.g., image catalogue, musical dictionary, biomedical imaging catalogues, film, video and radio archives)
- E-Commerce (e.g., personalised advertising, on-line catalogues, directories of e-shops)
- Education (e.g., repositories of multimedia courses, multimedia search for support material)
- Home Entertainment (e.g., systems for the management of personal multimedia collections, including manipulation of content, e.g. home video editing, searching a game, karaoke)
Applications domains (cont.)

- Investigation services (e.g., human characteristics recognition, forensics)
- Journalism (e.g. searching speeches of a certain politician using his name, his voice or his face)
- Multimedia directory services (e.g. yellow pages, Tourist information, Geographical information systems)
- Multimedia editing (e.g., personalised electronic news service, media authoring)
- Remote sensing (e.g., cartography, ecology, natural resources management)
- Shopping (e.g., searching for clothes that you like)
- Social (e.g. dating services)
- Surveillance (e.g., traffic control, surface transportation, non-destructive testing in hostile environments)
Organization of Multimedia DSs

**Content organization**
- Collection & classification
- Analytic model

**Creation & production**
- Media
  - Structural aspects
- Content management
- Conceptual aspects
- Usage

**Navigation and access**
- Summary
- Partition
- Variation

**User interaction**
- User preferences

**Basic elements**
- Datatype & structures
- Schema tools
- Link & media localization
- Basic DSs

**Conceptual aspects**

**Structural aspects**

**User preferences**
Audio descriptors
Low-level audio descriptors

- **Basic**: instantaneous waveform and power values
- **Basic Spectral**: a log-frequency power spectrum, and spectral features including spectral centroid, spectral spread, and spectral flatness
- **Signal parameters**: fundamental frequency of quasi-periodic signals, and harmonicity of signals
- **Timbral Temporal**: log attack time and temporal centroid
- **Timbral Spectral**: specialized spectral features in a linear-frequency space, including a spectral centroid, and spectral features specific to the harmonic portions of signals, including harmonic spectral centroid, spectral deviation, spectral spread, and spectral variation.
- **Spectral Basis representations**: features used primarily for sound recognition, but generally useful as projections into a low-dimensional space to aid compactness and recognition.
Low-level audio descriptors

**Basic**: instantaneous waveform and power values
Low-level audio descriptors

- **Basic**: instantaneous waveform and power values
- **Basic Spectral**: a log-frequency power spectrum, and spectral features including spectral centroid, spectral spread, and spectral flatness

AudioSpectrumEnvelope description of a pop song.

The required data storage is $N \times M$ values where $N$ is the number of spectrum bins and $M$ is the number of time points.
Low-level audio descriptors

- **Basic**: instantaneous waveform and power values
- **Basic Spectral**: a log-frequency power spectrum, and spectral features including spectral centroid, spectral spread, and spectral flatness
- **Signal parameters**: fundamental frequency of quasi-periodic signals, and harmonicity of signals
Low-level audio descriptors

- **Basic**: instantaneous waveform and power values
- **Basic Spectral**: a log-frequency power spectrum, and spectral features including spectral centroid, spectral spread, and spectral flatness
- **Signal parameters**: fundamental frequency of quasi-periodic signals, and harmonicity of signals
- **Timbral Temporal**: log attack time and temporal centroid
Low-level audio descriptors

- **Basic**: instantaneous waveform and power values
- **Basic Spectral**: a log-frequency power spectrum, and spectral features including spectral centroid, spectral spread, and spectral flatness
- **Signal parameters**: fundamental frequency of quasi-periodic signals, and harmonicity of signals
- **Timbral Temporal**: log attack time and temporal centroid
- **Timbral Spectral**: specialized spectral features in a linear-frequency space, including a spectral centroid, and spectral features specific to the harmonic portions of signals, including harmonic spectral centroid, spectral deviation, spectral spread, and spectral variation.
Low-level audio descriptors

- **Basic**: instantaneous waveform and power values
- **Basic Spectral**: a log-frequency power spectrum, and spectral features including spectral centroid, spectral spread, and spectral flatness
- **Signal parameters**: fundamental frequency of quasi-periodic signals, and harmonicity
- **Timbral Temporal**: log attack time and temporal centroid
- **Timbral Spectral**: specialized spectral features in a linear-frequency space, including a spectral centroid, and spectral features specific to the harmonic portions of signals, including harmonic spectral centroid, spectral deviation, spectral spread, and spectral variation.

**Spectral Basis representations**: features used primarily for sound recognition, but generally useful as projections into a low-dimensional space to aid compactness and recognition.

**Silence**

A 10-basis component reconstruction showing most of the detail of the original spectrogram including guitar, bass guitar, hi-hat and organ notes. The left vectors are an AudioSpectrumBasis Descriptor and the top vectors are the corresponding AudioSpectrumProjection Descriptor. The required data storage is $10(M+N)$ values.
High-level audio description tools (Ds and DSs)

- **Musical Instrument Timbre** description tools
  - The descriptors relate to notions such as “attack”, “brightness” or “richness” of a sound.
  - The timbre descriptor for sustained harmonic sounds combines the above Timbral Spectral low-level descriptors with a log attack time descriptor.
  - The percussive instrument descriptor combines the Timbral Temporal low-level descriptors with a spectral centroid descriptor.
  - Comparisons between descriptions using either set of descriptors are done with an experimentally-derived scaled distance metric.

- **Musical Instrument Timbre Description Tool**
  - HarmonicInstrumentTimbre (LAT + timbre spectral)
  - PercussiveInstrumentTimbre (timbre temporal + SpectralCentroid)
High-level audio description tools (Ds and DSs)

Sound recognition tools
- a collection of tools for indexing and categorization of general sounds, with immediate application to sound effects.
  - Support for automatic sound identification and indexing
  - tools for specifying a taxonomy of sound classes
  - tools for specifying an ontology of sound recognizers.

The recognition tools use
- the low-level Spectral Basis descriptors as their foundation.
- These basis functions are then further segmented into a series of states that comprise a statistical model, such as a hidden Markov or Gaussian mixture model, which is then trained on the likely transitions between these states.
- This model may stand on its own as a representation, have a label associated with it according to the semantics of the original sound, and/or associated with other models in order to categorize novel sounds input into a recognition system.
Spoken Content description tools

- allow detailed description of words spoken within an audio stream.
- sacrifice some compactness for robustness of search.
- represent the output and what might normally be seen as intermediate results of Automatic Speech Recognition (ASR).

The tools can be used for two broad classes of retrieval scenario:
- indexing into and retrieval of an audio stream,
- indexing of multimedia objects annotated with speech

are divided into two broad functional units:
- the lattice, which represents the actual decoding produced by an ASR engine,
- the header, which contains information about the speakers being recognized and the recognizer itself.
MPEG-7 Audio: High Level DSs

- **SpokenContentDescription Tools**
  - Output of ASR
    - phone lattice or word lattice
    - spoken content DS stores these lattices instead of plain text
    - lattices are good for retrieval
  - **SpokenContentLattice**
    - representing the actual decoding produced by an ASR engine
  - **SpokenContentHeader**
    - contains information about the speakers being recognized and the recognizer itself

A lattice structure for an hypothetical (combined phone and word) decoding of the expression “Taj Mahal drawing …”.
Melody description tools

- **Melody description tools**: The Melody Contour DS
  - is a compact representation for melodic information, which allows for efficient and robust melodic similarity matching, for example, in query-by-humming.

- **MelodyContour Description Scheme**
  - 5-step contour representation
  - basic rhythmic information representation

- **MelodySequence Description Scheme**
  - supporting an expanded descriptor set and high precision of interval encoding
MPEG-7 Audio: High Level DSs

- **General Sound Recognition and Indexing Description Tool**
  - **SoundModel (SM) DS**
    - statistical model, such as HMM or GMM
    - SoundModelStatePath Descriptor
      - consists of a state sequence generated by a SM
    - SoundModelStateHistogram Descriptor
      - consists of a normalized histogram of the state sequence generated by a SM given an audio segment

- **SoundClassificationModel DS**
  - a trainable multi-way classifier based on SMs
    - speech vs. music, male vs. female, trumpet vs. violin
    - genre classification, voice recognition
Multimedia DS
MPEG-7: MMDS Basic Elements

- **Root, Top-level elements, Packages**
- **Datatype & structures**
- **Schema tools**
- **Time, Duration, Medialocators**
- **Basic DSs**
- **Language Annotation, Person, Place**
- **Link & media localization**
MPEG 7: Content Management and Description

**Content management**
- Creation & production
- Structural aspects
- Conceptual aspects
- Media

**Content description**
- Title, Creator, Creation location & date, Purpose, Classification, Genre, Review, Parental guidance, etc. (Author generated)
- Format, Coding, Instances, Identification, Transcoding Hint, etc. (Several instances)
- Rights holder, Access rights, Usage Record, Financial aspects, etc. (Evolution)

**Viewpoint of the structure: Segments**
- Spatial / temporal structure
- Audio, video low-level Ds
- Elementary semantic information.

**Viewpoint of conceptual notions**
- Events, objects, abstract concepts, and their relation
Example of decomposition and region-based description

Example of image description with Still Regions (SR)

- **SR1:** Creation, usage metainformation, Media description, Textual annotation, Color histogram, texture
- **SR2:** Shape, Color histogram, Textual annotation
- **SR3:** Shape, Color histogram, Textual annotation
- **SR4:** Shape, Color histogram, Textual annotation
- **SR5:** Shape, Textual annotation
- **SR6:** Color histogram, Textual annotation
- **SR7:** Color histogram, Textual annotation
Example of decomposition and region-based description

Example of image description with Still Regions (SR)

**SR1:**
- Creation\Usage meta information
- Media description
- Textual annotation
- Color histogram, Texture

**SR2:**
- Shape
- Color histogram
- Textual annotation

**SR3:**
- Shape
- Color histogram
- Textual annotation

**SR4:**
- Shape
- Color histogram
- Textual annotation

**SR5:**
- Shape
- Textual annotation

**SR6:**
- Color histogram
- Textual annotation

**Foreground**

**Background**
Segment Tree

Segment 1
- Sub-segment 1
- Sub-segment 2
- Sub-segment 3
- Sub-segment 4

Segment 2

Segment 3

Segment 4

Segment 5

Segment 6

Segment 7

Semantic DS (Events)

- Introduction
- Summary
- Program logo

- Studio
  - Overview
  - News Presenter

- News Items
  - International
    - Clinton Case
    - Pope in Cuba
  - National
    - Twins

- Sports

- Closing
Example of decomposition and region-based description

Example of video segments and regions

- two Video segments, one Still Region and three Moving Regions
Example

Segment Relationship Graph

Video segment 1: Pass

Video segment 2: Kick and score

Graph describing the structure of the content
Efficient support of: discovery, browsing, navigation, visualization

Substitution of the original content
Adaptation to terminal, network, or user preferences
MPEG 7: Hierarchical summary

A-V Data

Hierarchical Summary

Highlight Level

Highlight Segment

Highlight Segment

Highlight Segment

Highlight Segment

Highlight Segment

Highlight Segment

Highlight Segment
MPEG 7: Sequential summary

Sequential Summary

- Frame Property
- Text Property
- Sound Property

A-V Data

Frame Property

Text Property

Sound Property

Frame Property

Sound Property
MPEG 7: Variation

Universal Multimedia Access

- Adapt delivery to network and terminal characteristics (QoS)
MPEG-7: Content Organization

**Content organization**

- Description and organization of collection of documents

**Content management**

- Media
- Structural aspects

**Content description**

- Content Usage
- Conceptual aspects

**Basic elements**

- Datatype & structures
- Schema tools
- Link & media localization
- Basic DSs

**Collection & Classification**

**Analytic Model**

**Navigation & Access**

- Summary
- Variation
MPEG-7: Collection

Collection Structure

Collection A

Collection B

Collection C

R_AB

R_BC

R_AC
MPEG 7: User Interaction

### Content organization

#### Media

- Content management
  - Creation & production
  - Content description
  - Structural aspects
  - Conceptual aspects

#### Collection & Classification

- Analytic Model

#### Navigation & Access

- Summary
- Variation

### Basic elements

- Datatype & structures
- Schema tools
- Link & media localization
- Basic DSs

**User identification and preferences:**
Filtering, search and browsing