7 Multimedia Content Description

Literature:

Rosenblatt/Trippe/Mooney, Digital Rights Management, Chapter 6

Unlabelled Video Tapes & The Internet

- The Unlabeled Video Tape Problem
  - Even worse with digital media: Various formats, variants
- Digital media production:
  - Labeling of parts to be composed
    - » Date, time, format, ...
  - Representing the composition
- Digital media on the Internet
  - Identifying digital media
    - » Title, author, genre, ...
  - Searching for specific media, e.g. audio, video content
  - Fine-grained search within media
    - » e.g. person search within video content
  - Bringing together related media (e.g. text news and photos)
    - » (Automated) syndication

Content, Essence, Metadata

- Content
  - consists of *essence* data and *metadata*
- Essence
  - parts of content that directly represent program material such as audio, video, graphic, still-image, text, or sensor-data
- Metadata
  - parts of content that contain data used
    - » to *describe* essence or
    - » to provide information on its *use*
  - metadata objects sometimes called “mobs”
- Metadata may be
  - Stored separately from the essence data
  - Combined with the essence data (“embedded metadata”)

Source: AAF Developer Overview

Types of Multimedia Metadata

- Technical Metadata:
  - Form (data format, representation parameters like resolution, color depth...)
  - For live captured material: Time, date, location of original occurrence
  - Technical parameters of capture (e.g. aperture, exposure etc. for images)
- Content Description Metadata:
  - High level, structured:
    - » Title, author, composer, interpret, cast, ...
  - High level, unstructured:
    - » Summary, textual description, thumbnail, ...
  - Low level:
    - » Objects and time positions
    - » Audio and video attributes: Key, mood, tempo ...
- Additional information:
  - Digital rights, classification, context, further links, ...

Types of Origin for Metadata

- Automatic creation or derivation:
  - All technical metadata
  - Some low level metadata (e.g. average brightness, musical tempo)
- Retrieval from external databases:
  - High-level metadata
  - Retrieval may be based on identifier or analysis of media content
  - Example: GraceNote database for music
- Manual addition
  - Archival, indexing, annotation, ...

Metadata Problems

- Creation metadata
  - During the creation of media essence, metadata is created but often ignored
  - Example: EXIF data in JPEG
- Manually added metadata
  - Users notoriously ignore the administration of metadata
- Metadata incompatibility
  - Metadata exists in various formats specific for media types, applications, product vendors, ...
  - Exchange of metadata is difficult
- Broad range of metadata
  - Metadata exists on various levels, covering all is expensive
- Metadata economy
  - How much of the metadata will be used?
  - When to create metadata?

7 Multimedia Content Description

Literature:

B. S. Manjunath, Philippe Salembier, Thomas Sikora:
Introduction to MPEG-7, Wiley 2002

www.chiariglione.org
mpeg-7.joanneum.at
www.multimedia-metadata.info

MPEG-7

- Moving Picture Experts Group (MPEG)
  - = ISO/IEC JTC1/SC29/WG11 “Moving Pictures and Audio”
  - Main Web presence now: www.chiariglione.org/mpeg
- MPEG-7 “Multimedia Content Description Interface”
  - “ ... a standard for describing the multimedia content data that supports some degree of interpretation of the information’s meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 is not aimed at any one application in particular; rather, the elements that MPEG-7 standardizes support as broad a range of applications as possible.”
- Version 1 developed in 1996 – 2001, standard since 2002
- Industrial uptake very slow
  - Ambitious standard
- Some research and open source prototypes available
  - See e.g.
    http://mpeg7.joanneum.at,
    http://www.multimedia-metadata.info

Parts of the MPEG-7 Standard

- MPEG-7 Systems
- MPEG-7 Description Definition Language (DDL)
  - Descriptors (D) define the syntax and semantics of each *feature* (metadata element)
  - Description schemes (DS) specify syntax and semantics of the relationships between their components, which may be Descriptors or Description Schemes
  - DDL allows the creation of Ds and DSs
    - » XML-based language with some small extensions to XML Schema
- MPEG-7 Visual
- MPEG-7 Audio
- MPEG-7 Multimedia Description Schemes
- MPEG-7 Reference Software
  - eXperimentation Model XM

Application Areas of MPEG-7

- Architecture, real estate, and interior design (e.g., searching for ideas).
- Broadcast media selection (e.g., radio channel, TV channel).
- Cultural services (e.g., virtual museums).
- Digital libraries (e.g., image catalogue, musical dictionary).
- Education (e.g., repositories of multimedia courses).
- Home Entertainment (e.g., home video management).
- Investigation services (e.g., human characteristics recognition, forensics).
- Journalism (e.g. searching for video footage of political event).
- Multimedia directory services (e.g. yellow pages, Tourist information).
- Multimedia editing (e.g., personalized electronic news service, media authoring).
- Remote sensing (e.g., cartography, ecology, natural resources management).
- Shopping (e.g., searching for clothes that you like).
- Surveillance (e.g., traffic control, surface transportation).
- ...

Examples of Advanced Queries

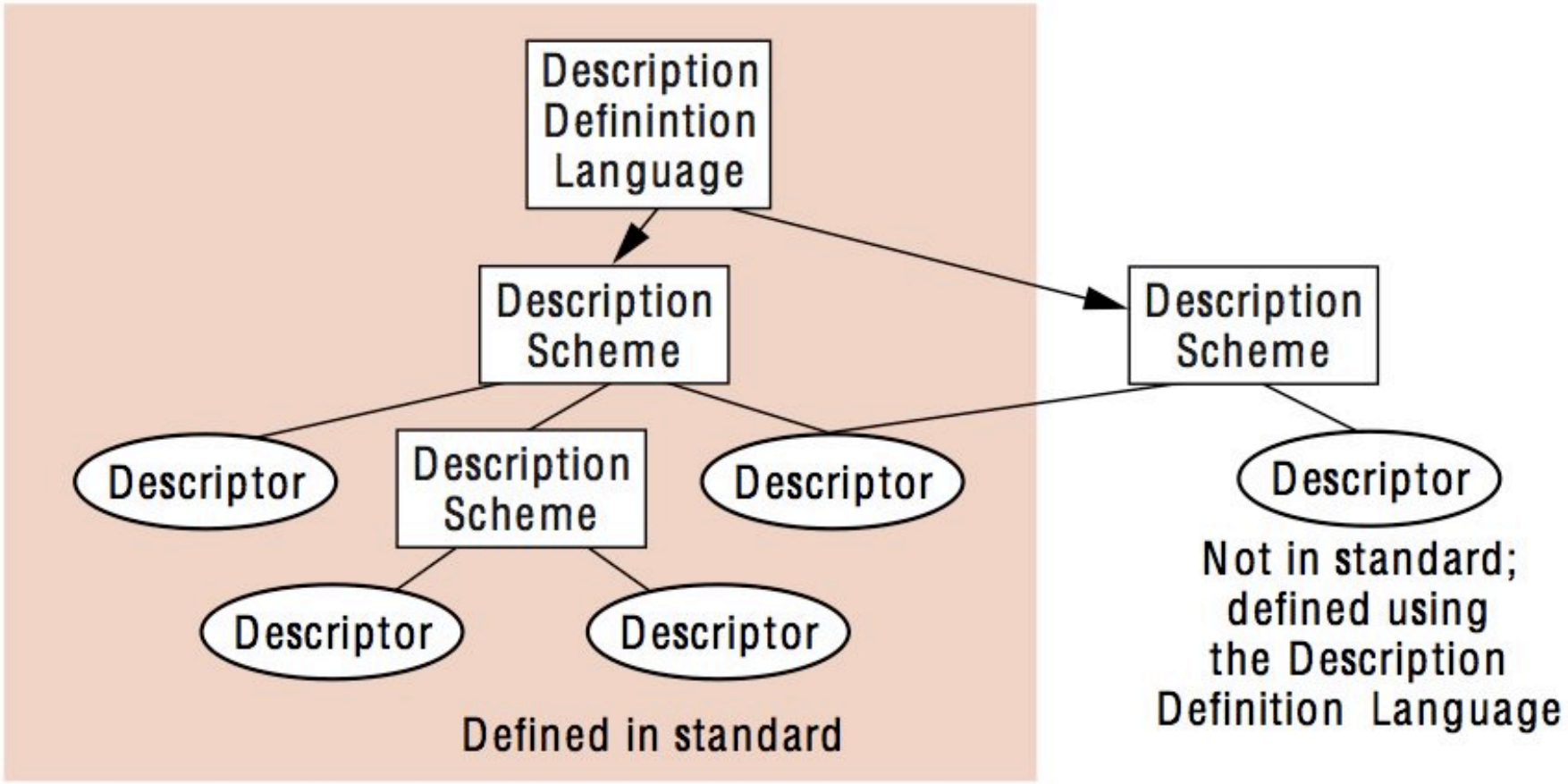
- Play a few notes on a keyboard and retrieve a list of musical pieces similar to the required tune, or images matching the notes in a certain way, e.g. in terms of emotions.
- Draw a few lines on a screen and find a set of images containing similar graphics, logos, ideograms,...
- Define objects, including color patches or textures and retrieve examples among which you select the interesting objects to compose your design.
- On a given set of multimedia objects, describe movements and relations between objects and so search for animations fulfilling the described temporal and spatial relations.
- Describe actions and get a list of scenarios containing such actions.
- Using an excerpt of Pavarotti's voice, obtaining a list of Pavarotti's records, video clips where Pavarotti is singing and photographic material portraying Pavarotti.

From: MPEG-7 Overview

MPEG-7 Description Terminology (1)

- Feature:
  - Distinctive characteristic of the data which signifies something to somebody
- Descriptor:
  - Representation of a feature
    - » Defines syntax and semantics of feature representations
  - A feature may be represented by several descriptors
- Descriptor value:
  - Instantiation of a descriptor
- Description scheme:
  - Structured composition of descriptions and description schemes
- Description:
  - Instance of a description scheme with appropriate descriptor values

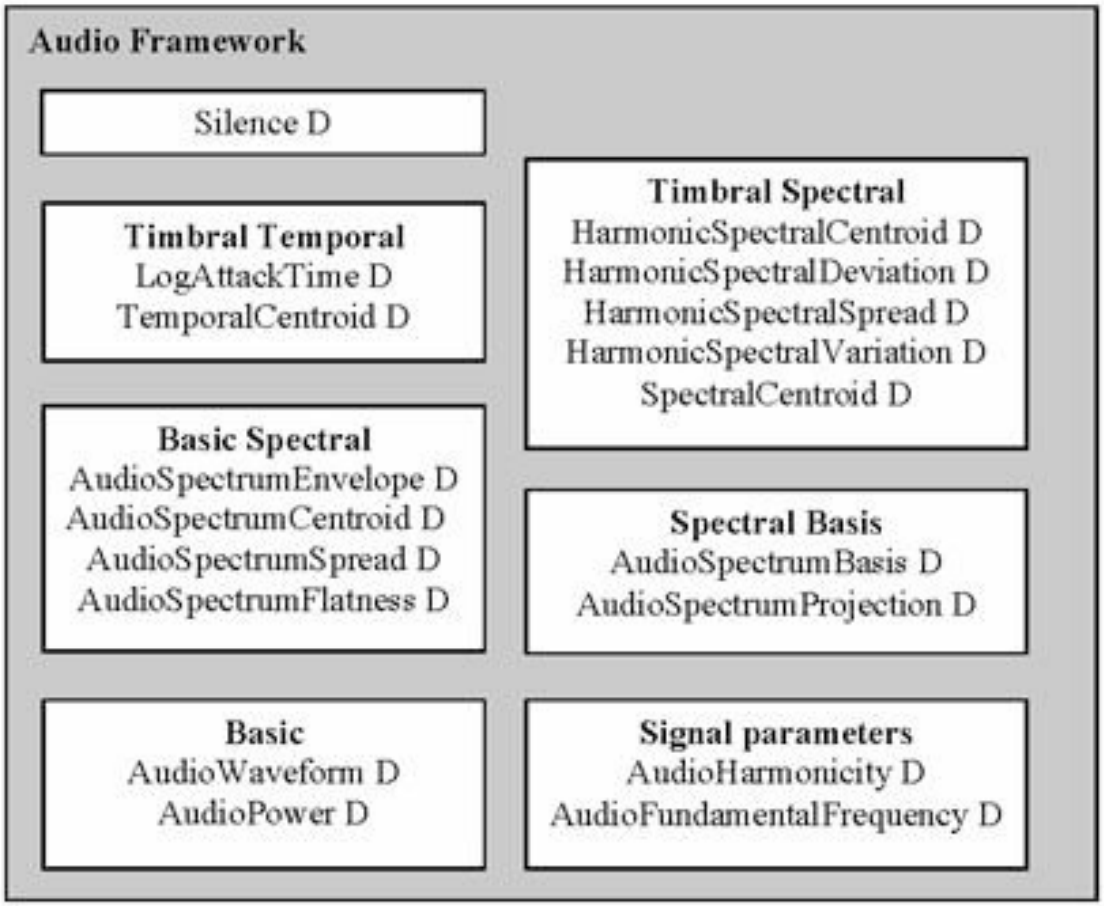
MPEG-7 Description Terminology (2)



Nack/Lindsay: Everything You Wanted to Know About MPEG-7, Part 2, *IEEE Multimedia* Magazine, October 1999

MPEG-7 Audio Low-Level Descriptors

- Structures:
  - Single scalar value
  - Series of sampled values
- Features:
  - See figure



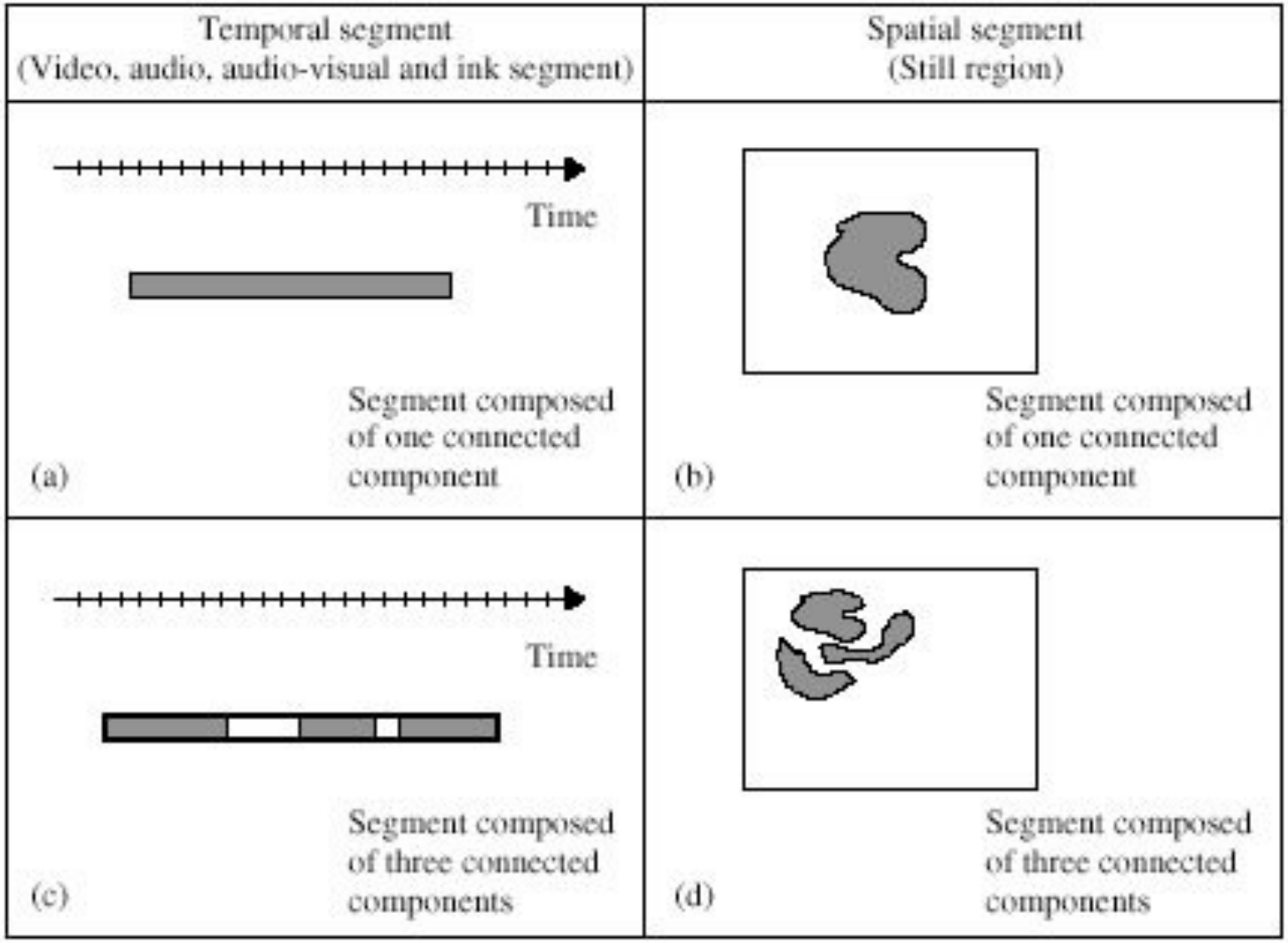
MPEG-7 Audio High-Level Descriptors

- Audio signature
  - Statistical summary of spectral flatness descriptor
  - Fingerprinting, identification of audio content
- Musical instrument timbre
- Melody description
  - MelodyContour (terse, efficient)
  - MelodySequence
    - » Query by Humming
    - » Example: http://www.musicline.de/de/melodiesuche
- General sound recognition and indexing
  - Probabilistic classifiers for sound classes
- Spoken content
  - Output and intermediate results of Automatic Speech Recognition (ASR)

Structural Content Description: Segments

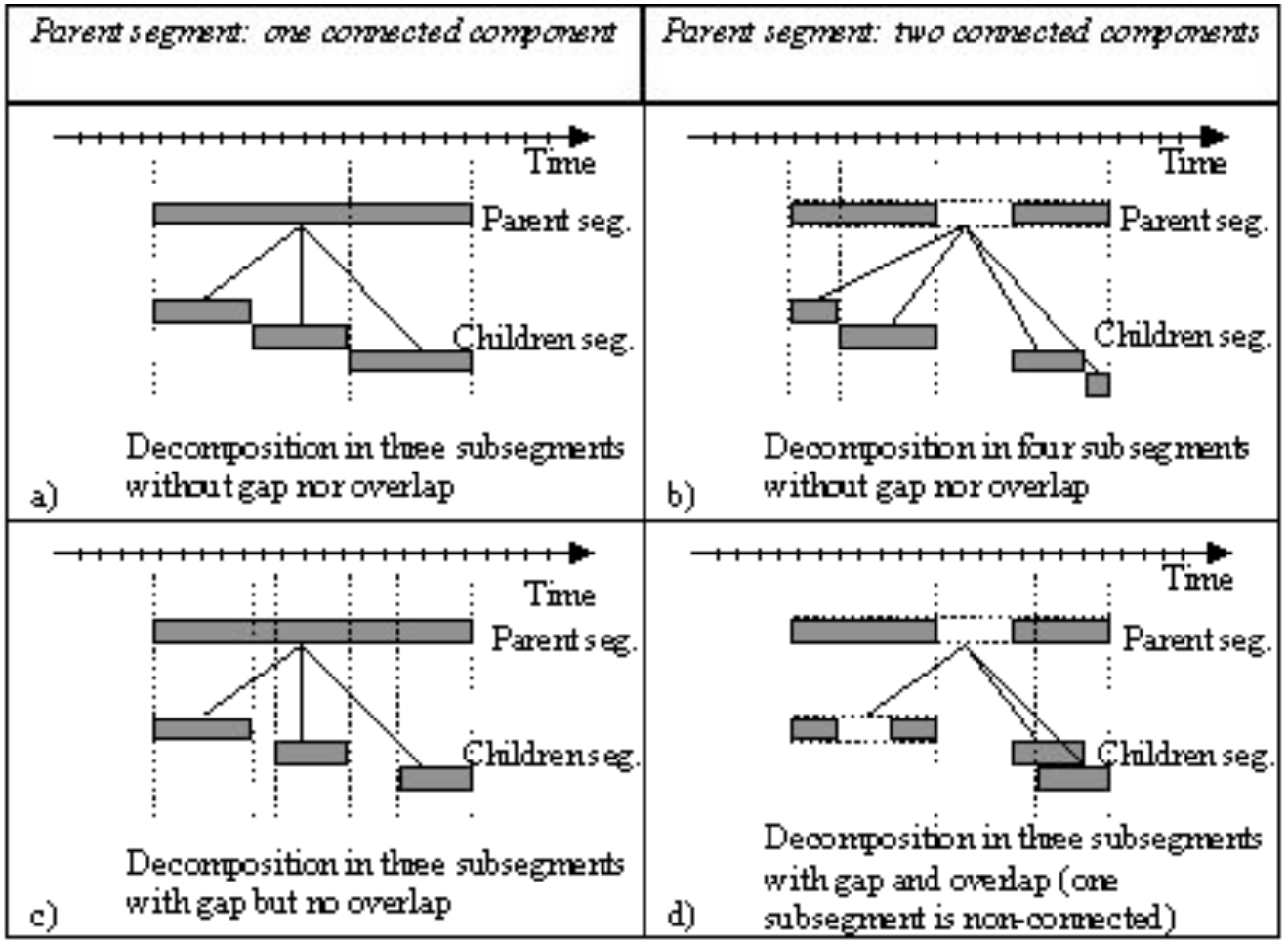
- A segment represents a section of an audio-visual content item.
- The Segment Description Scheme (DS) is an abstract class (in the sense of object-oriented programming).
- It has nine major subclasses:
  - Still Region DS (spatial)
  - Video Segment DS (temporal)
  - Moving Region DS (spatiotemporal)
  - Audio Segment DS (temporal)
  - AudioVisual Segment DS (temporal)
  - AudioVisual Region DS (spatiotemporal)
  - Still Region 3D DS (3D spatial)
  - Ink Segment DS (electronic ink from pen, smartboard etc.)
  - Multimedia Segment DS (composite of segments)

Examples of Segments

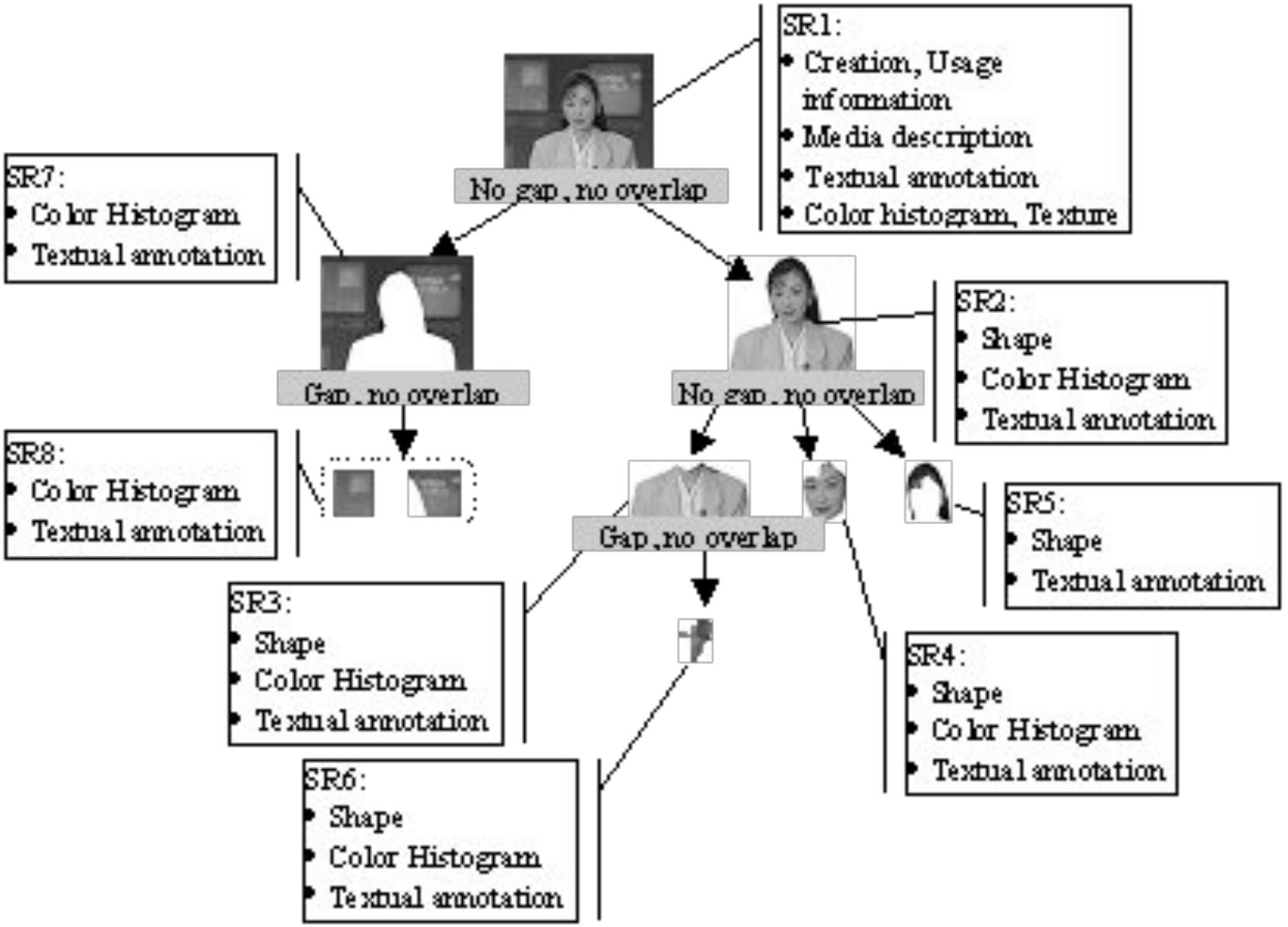


Segment Decomposition

- Segments can be decomposed into subsegments
  - Subsegments may overlap in time/space
  - Subsegments may not cover the full extents of parent segment
  - Decomposition may result in segments of different nature



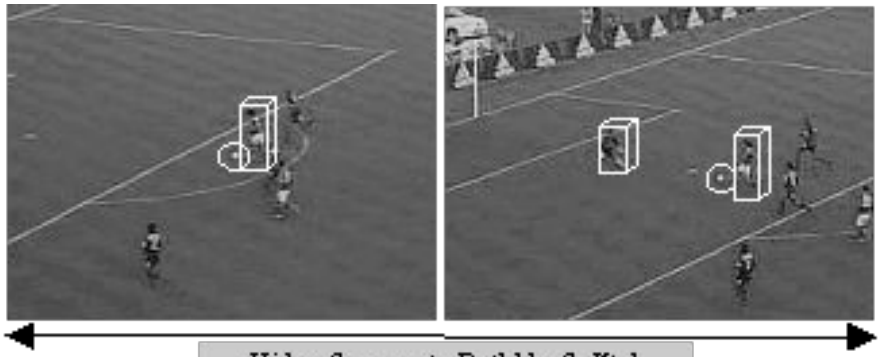
Example of Image Description



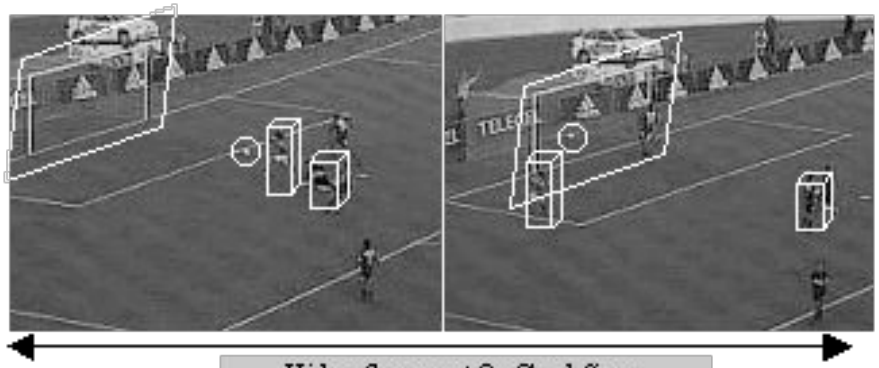
Structural Relations of Segments

- Content structure:
  - Either hierarchical segment decomposition
  - Or general segment relationship graph
- Predefined structural relations in MPEG-7 (can be extended):
  - Generic:
    - » Identical, union, disjoint
  - Spatial:
    - » South, north, west, east, northwest, northeast, southwest, southeast, left, right, below, above, over, under
  - Temporal:
    - » Precedes, follows, meets, metBy, overlaps, overlappedBy, contains, during, strictContains, strictDuring, starts, startedBy, finishes, finishedBy, coOccurs, contiguous, sequential, coBegin, coEnd, parallel, overlapping
- For each relation, the inverse relation is implicitly defined.

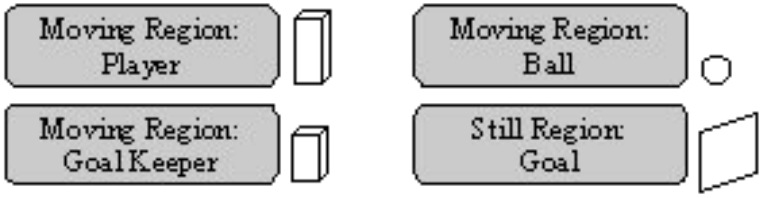
Video Segmentation with Moving Regions

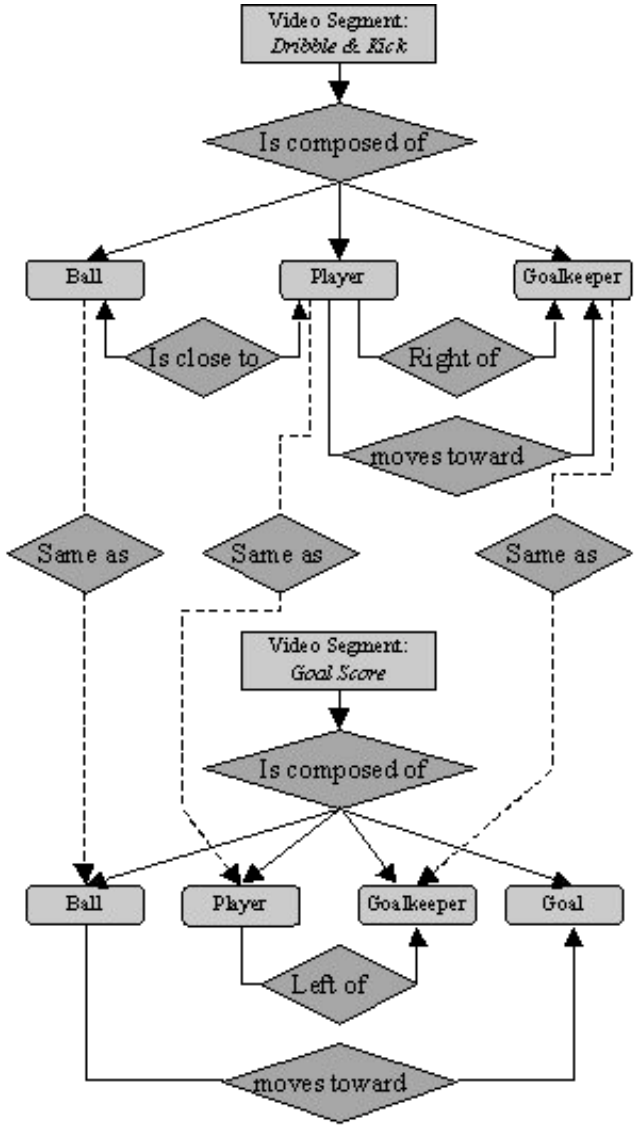


Video Segment: *Dribble & Kick*



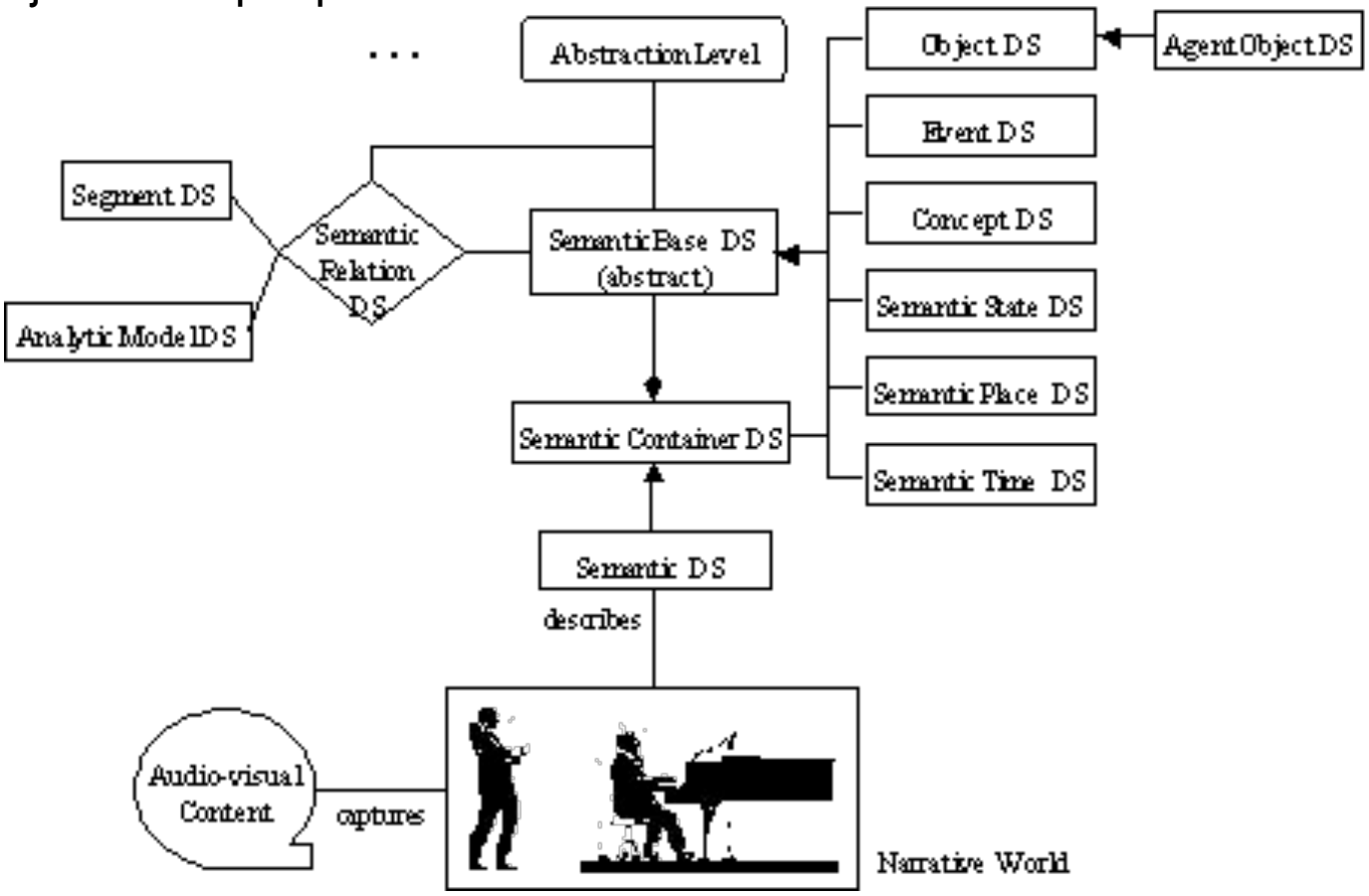
Video Segment 2: *Goal Score*



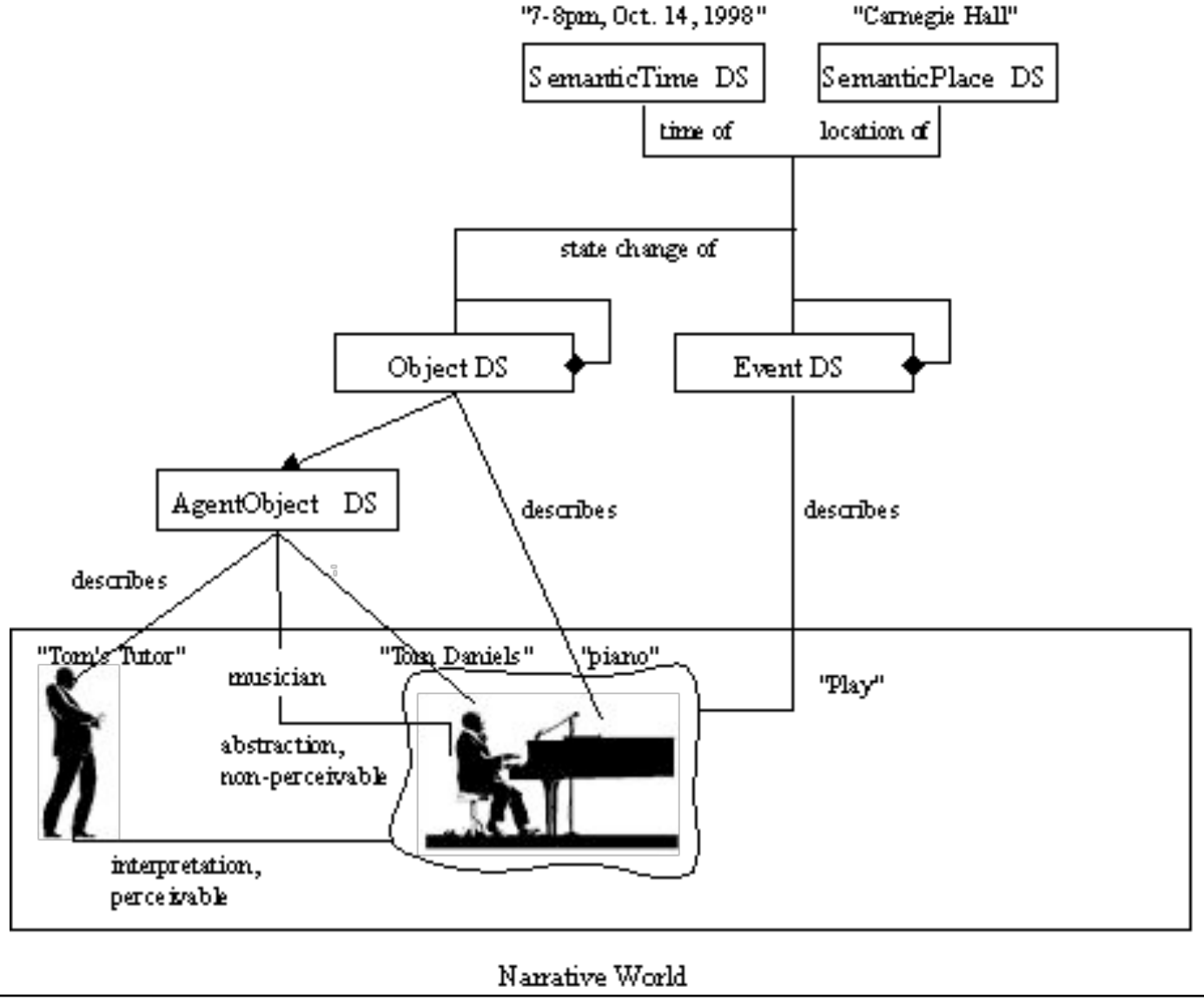


Content Semantics in MPEG-7

- Event: Occasion when something happens
  - Occurs at some time and place
  - Populated by objects and people
- "Narrative world" for a piece of content

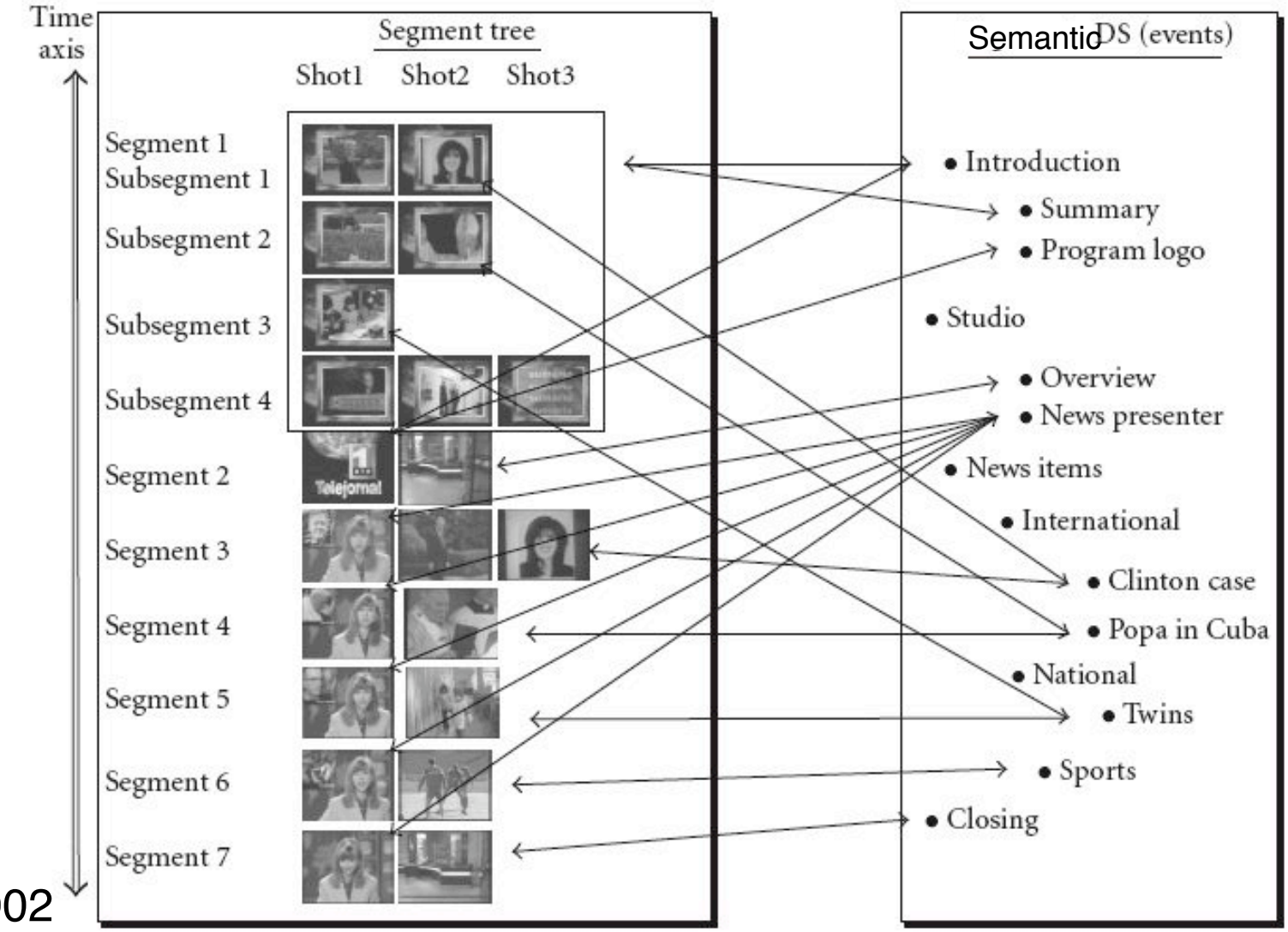


Content Semantics in MPEG-7: Example



Relating Structure and Semantics: Example

- Creating a semantic index for a video sequence



Salembier 2002

7 Multimedia Content Description

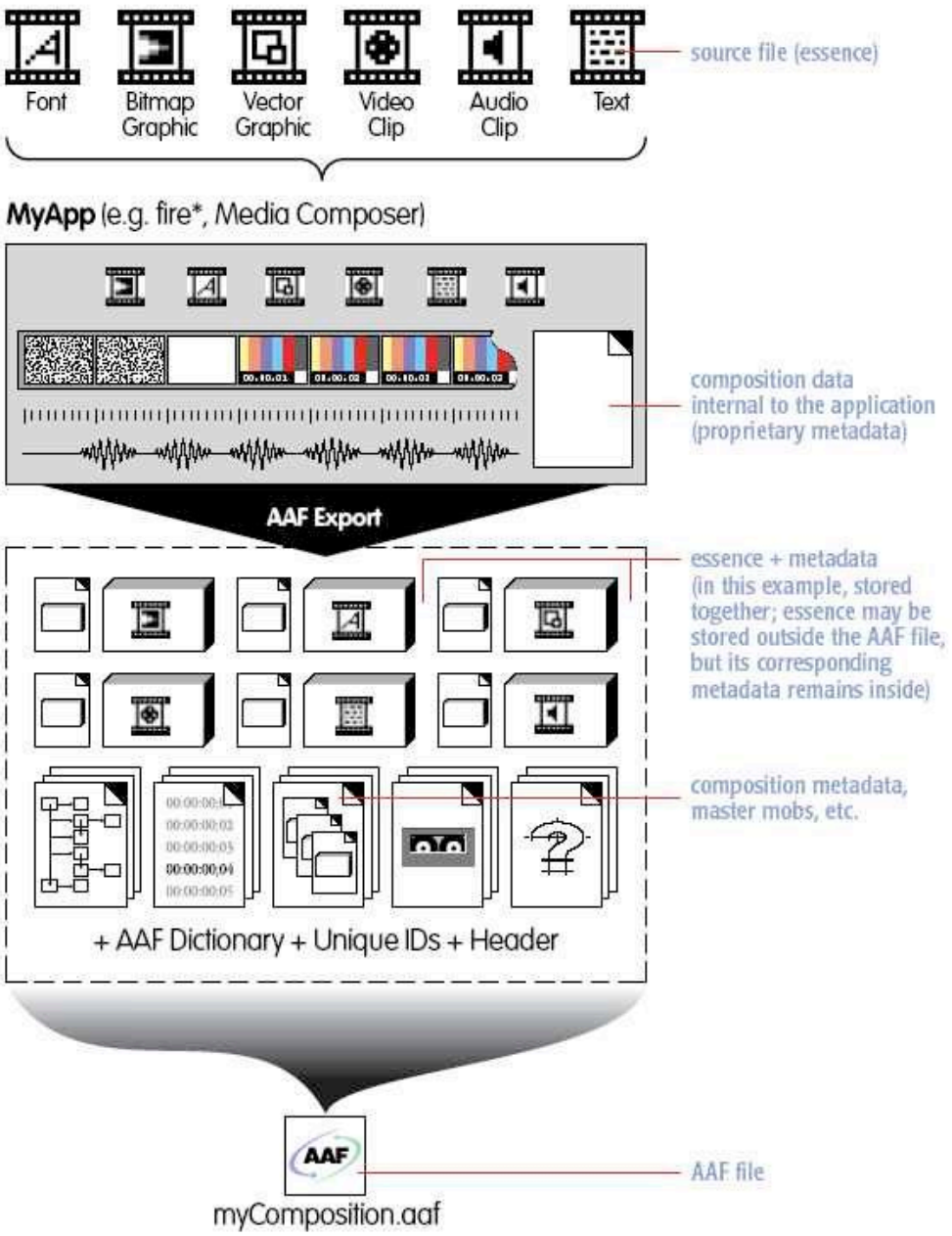
Selected Media Metadata Standards

- Dublin Core Metadata Initiative (DCMI) &
  PRISM (Publishing Requirements for Industry Standard Metadata)
  - Oriented towards books, magazines, journals etc.
  - Uses XML, RDF, Dublin Core
  - dublincore.org, www.prismstandard.org
- ONIX (Online Information Exchange)
  - For books: http://www.editeur.org/8/ONIX
- TV Anytime (www.tv-anytime.org)
  - Devoted to audio-visual services making use of local mass-storage
  - Focus on: Electronic Program Guide and user profiles
- EBU P/Meta
  - Devoted to material exchange between broadcasting stations
- Commercial solutions by Rovi (www.rovicorp.com), *ex* Macrovision
  - Company trying to set de-facto industry standard
  - Company collecting large database of media metadata (e.g. www.muze.com)

Integration of Digital Media in Video Production

- Example: Putting together all audio elements for a film soundtrack
  - Music tracks, ambient sound tracks, performer's synchronized sound, ...
  - Metadata related to creation process need homogeneous treatment
- Standards in the broadcasting industry
  - SMPTE (Society of Motion Picture and Television Engineers)
  - EBU (European Broadcasting Union)
  - Working on hardware-based standards for a long time
- EBU/SMPTE Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams (1996-1999)
  - Results further developed into Advanced Authoring Format (AAF)
  - AAF: Industry-driven, cross-platform, multimedia file format
  - "Advanced Media Workflow Association"
    - » see www.aafassociation.org

Interchanging Compositions with AAF



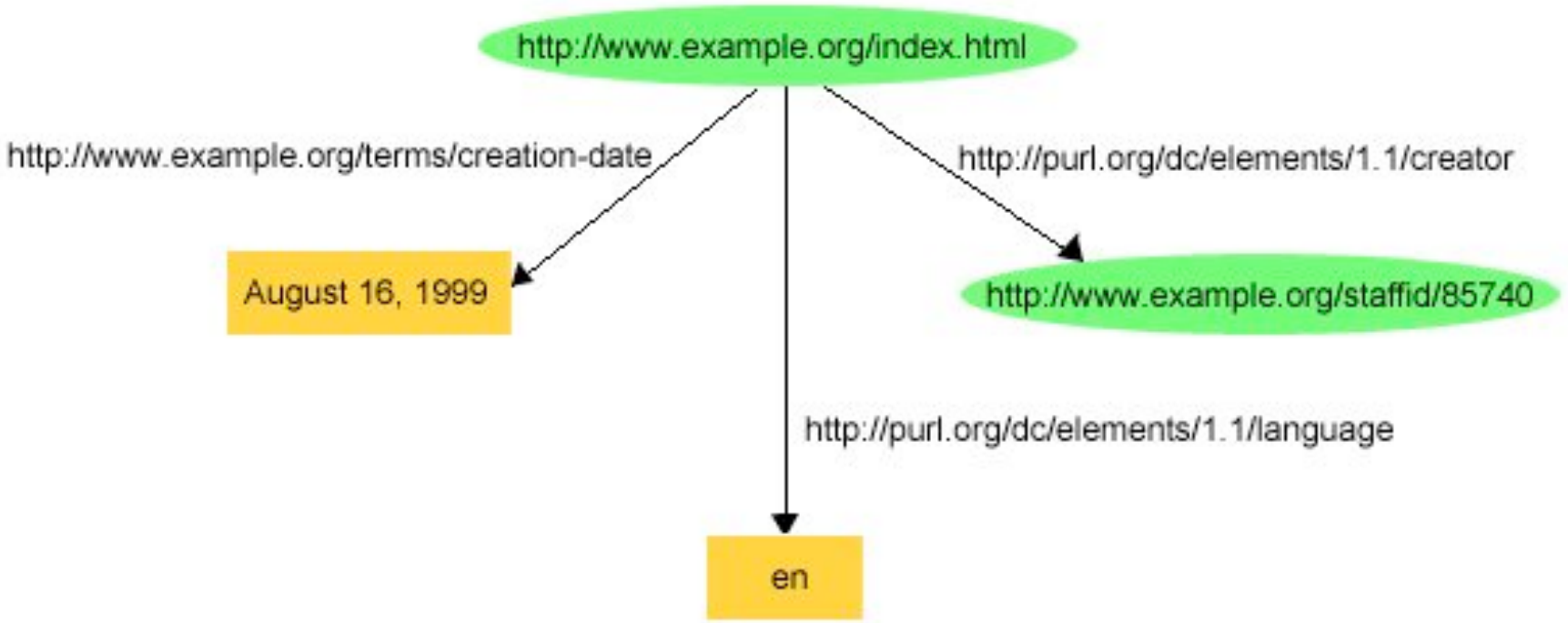
Resource Description Framework RDF

- Language for representing information about resources in the WWW
  - W3C's Semantic Web activity

- *Resource:* Anything that can be identified by a URI (e.g. all Web pages)
- *Property:* An attribute of a described thing which can take on specific values
- *Statement:* A triple consisting of
  - *Subject:* Some resource to be described
  - *Predicate:* A property of the subject
  - *Object:* A specified value
- Properties, values and statements are resources themselves,
  - i.e. can be identified by a URI
  - i.e. can be subject to further description

RDF Example

- **http://www.example.org/index.html** has a **creator** whose value is **John Smith**
- **http://www.example.org/index.html** has a **creation-date** whose value is **August 16, 1999**
- **http://www.example.org/index.html** has a **language** whose value is **English**



Example: Audio Metadata in DC-based RDF/XML

RDF/XML is an XML language for representing descriptions

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
     rdf:about="http:www.medien.ifi.lmu.de/team/
     heinrich.hussmann/files/mmn8a.m4b>
    <dc:creator>Heinrich Hussmann</dc:creator>
    <dc:title>Multimedia Content Description I</dc:title>
    <dc:description>Discusses multimedia metadata
     standards.</dc:description>
    <dc:date>2009-12-10</dc:date>
    <dc:format>audio/mp4</dc:format>
  </rdf:Description>
</rdf:RDF>
```

7 Multimedia Content Description

7.1 Metadata: Concepts and Overview

7.2 MPEG-7: Generic Metadata Standard

7.3 Selected Metadata-Relevant Standards

7.4 Automated Extraction of Multimedia Metadata

– Music, Images, Video

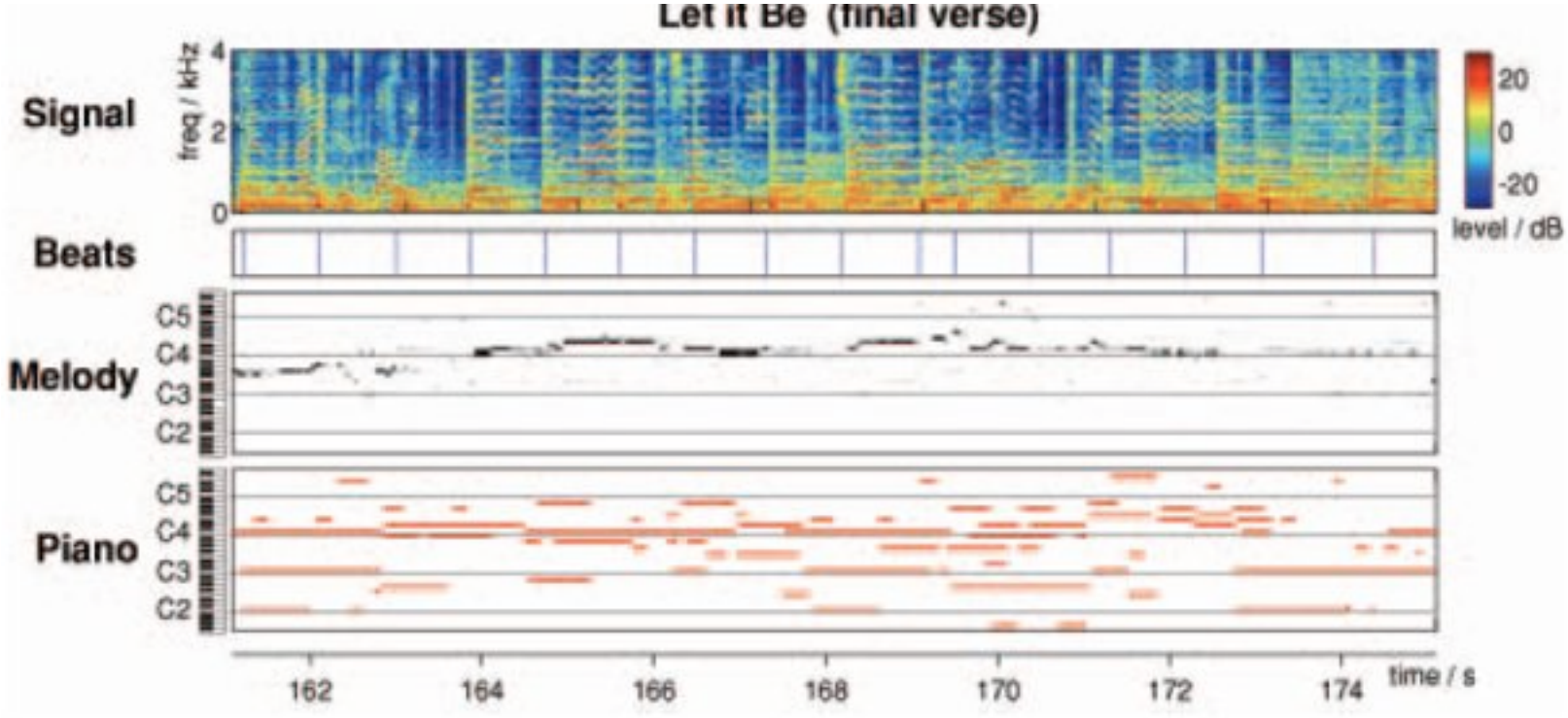
Literature:

Communications of the ACM 49(8), August 2006,
Special section on Music Information Retrieval, pp. 28-60

Timescales of Musical Information

- Individual music note events
  - Extraction of the music score
  - Identification of instrument playing
- Chords (simultaneous notes)
  - Identification of chords
- Phrase level
  - Tempo extraction
  - Identification of phrases (based on repetition/alternation of segments) e.g. identificaton of chorus
- Piece level
  - Genre identification (“rock”, “jazz”, “classical”)

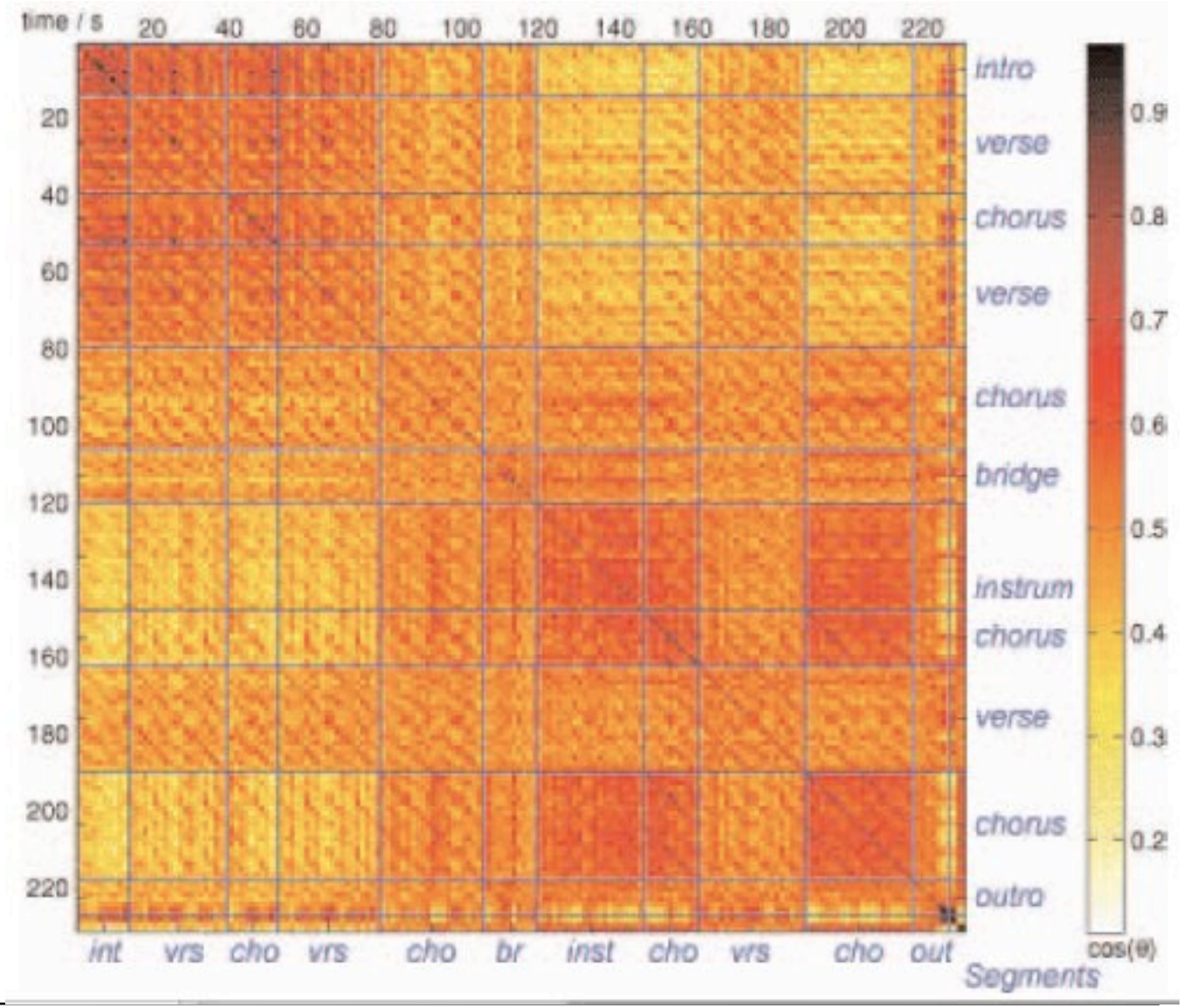
Automatic Score Transcription



- Beats determined by tempo-smoothed event detector
- Melody recognized by general-purpose support-vector classifier
  - Trained to recognize spectral slices to be labelled with pitch values

Automatic Phrase Detection

- Self-similarity matrix
  - Values represent acoustic similarity
  - Looking for diagonal ridges off the main diagonal
  - Blue lines are manually inserted for comparison



See also:
http://www.fxpal.com/publications/FXPAL-PR-99-093.pdf

Example: Shazam Music Tagging (1)

- Commercial service for mobile phones:
  Identify music from a short audio sample *(query by example)*
  - See http://www.shazam.com (London, founded 2000)
  - A. Wang: The Shazam Music Recognition Service, *Comm. ACM* Aug. 2006
- Challenges:
  - Distinguishing music from noise
  - Dealing with distortions
  - Keeping fingerprints small (in order to deal with millions of songs)
- Basic idea:
  - Spectrogram peaks (energy distribution in time and frequency)[1]
  - Few “anchor” peaks are combined with peaks in a certain surrounding zone (time and frequency offsets)
    - » Combinatorial hashing creates 32b fingerprint hash token

[1] An overlapping Short-Time Fourier Transform is calculated at regular intervals on the audio data, and a power level is calculated for each resulting time-frequency bin. A bin is a peak if its power level is greater than all the other bins in a bounded region around the bin.

Example: Shazam Music Tagging (2)

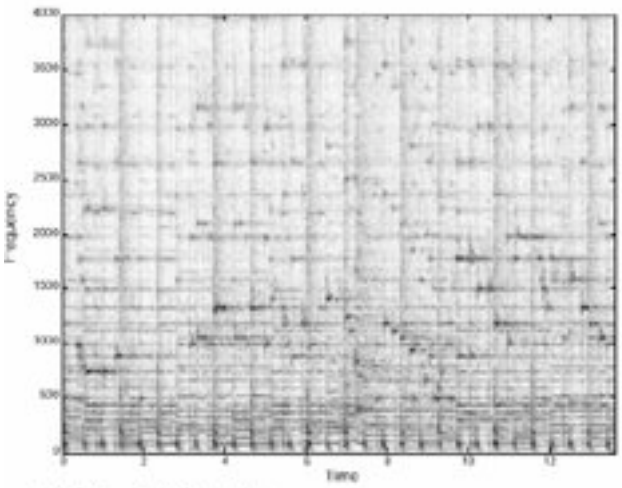


Fig. 1A - Spectrogram

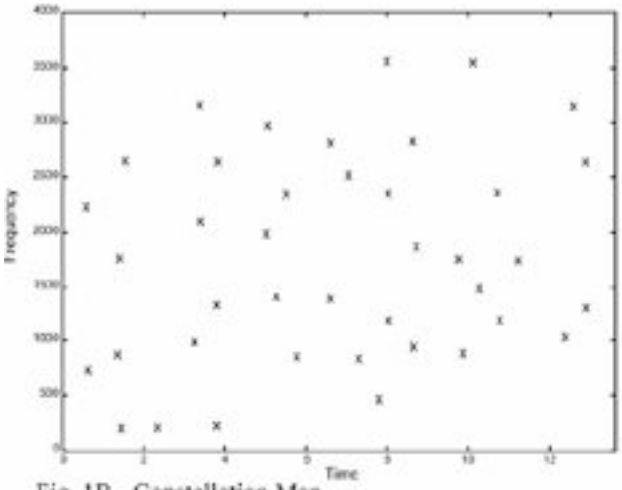


Fig. 1B - Constellation Map

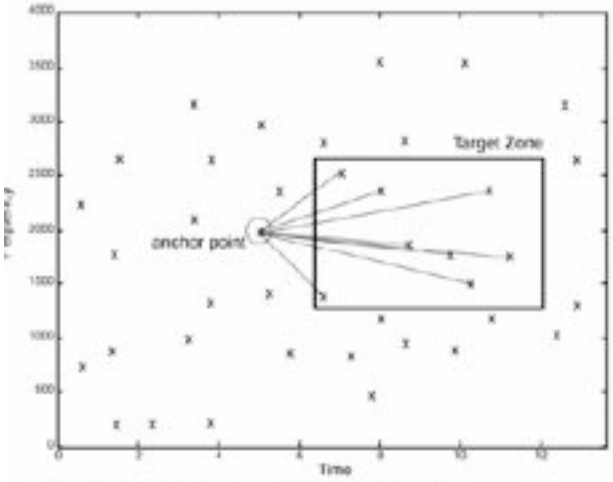


Fig. 1C - Combinatorial Hash Generation

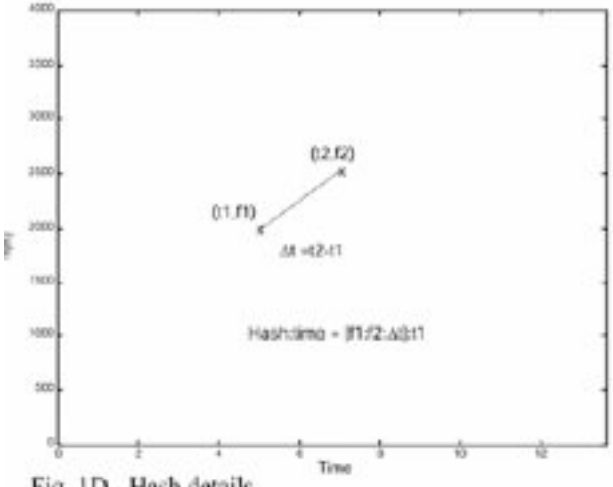


Fig. 1D - Hash details

http://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf

Fingerprint Complexity Tradeoff

- Computing a more complex fingerprint:
  - Increases search time (more tokens to inspect)
  - Improves entropy
    - » Better descriptiveness distinguishes more clearly between items
- Shazam example:
  - Combinatorial expansion increases token number by factor 10 (roughly)
  - Combinatorial expansion accelerates index search by a factor of more than a million!

Example: Shazam Music Tagging (3)

- Comparing tokens from sample and database:
  - Only tokens having peaks from target signal are relevant
  - Even presence of a few well matching tokens is significant
- Temporal alignment of fingerprint features:
  - Matching set of features must have identical relative positions in time
  - Find linear time correspondence
    - » By searching a histogram of relative time differences for peaks

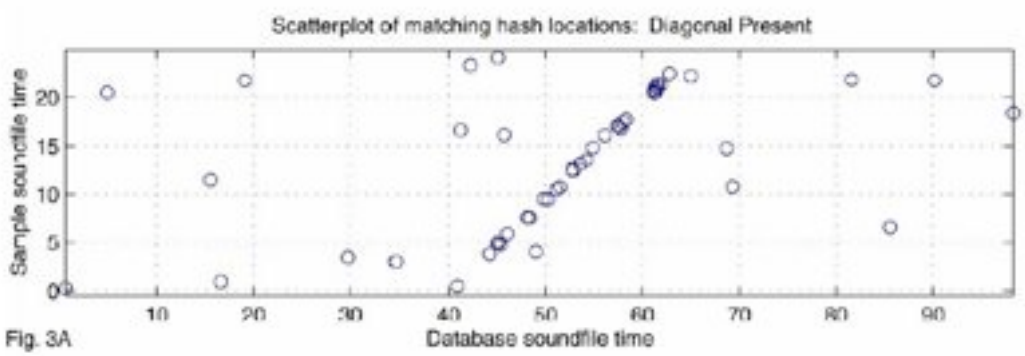


Fig. 3A

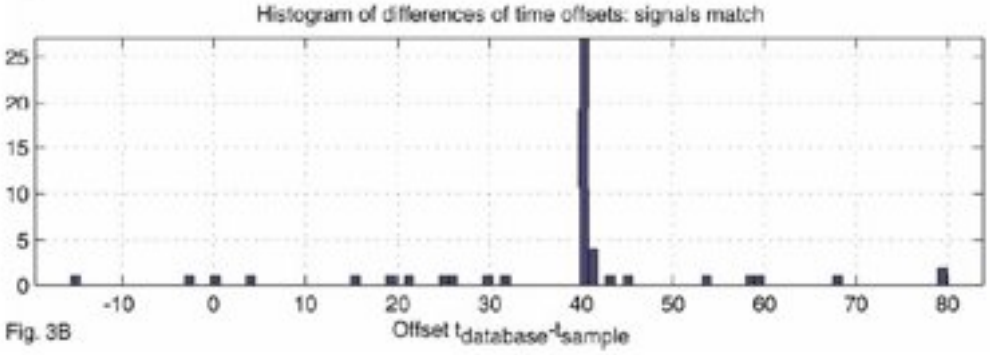


Fig. 3B

Example: Shazam Music Tagging (4)

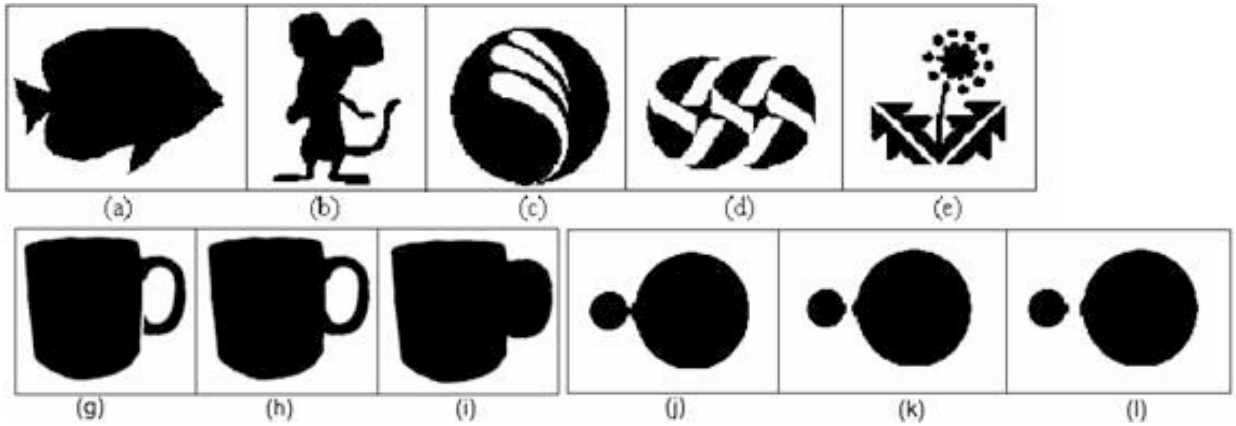
- Commercial situation:
  - 2009, more than 8 million tracks in database
  - By end of 2009, more than 250 million queries processed
- Without Internet connectivity:
  - Query via speech channel, result via text message
- (Free!) iPhone application:
  - Requires Internet connectivity
  - Query and result via Internet
  - Comfortable integration with other services (e.g. iTunes)

MPEG-7's Sophisticated Shape Descriptors

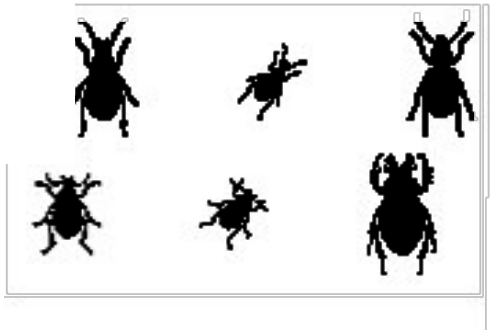
- Region shapes
  - Pixel distribution, using both boundary and internal pixels
  - Can describe complex objects with multiple disconnected regions
  - Shape analysis based on moments
    - » Angular Radial Transformation (ART)
- Contour shapes
  - Based on Curvature Scale-Space (CSS) representation of contour
  - Recognized characteristic contour shapes
  - Similar to human perception
- Desirable properties of extraction methods
  - Able to handle complex shapes
  - Robust to minor deformations, perspective transformations, movement, splits, occlusions etc.
  - Compact and efficient

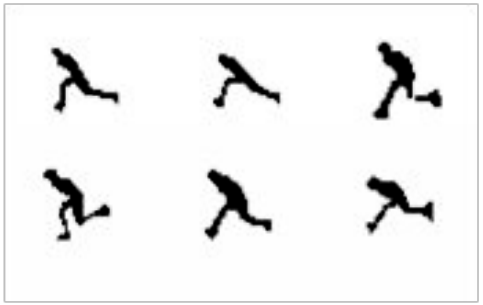
Examples for Shape Descriptors

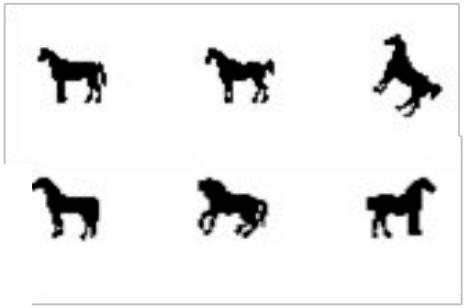
Region shapes:



Contour shapes:

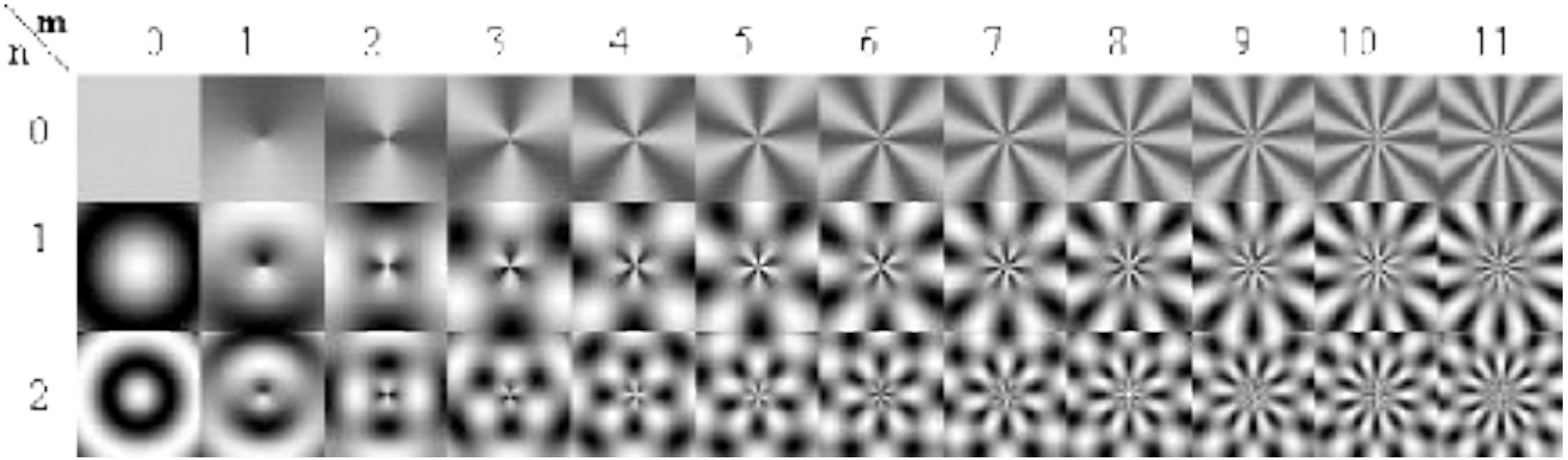






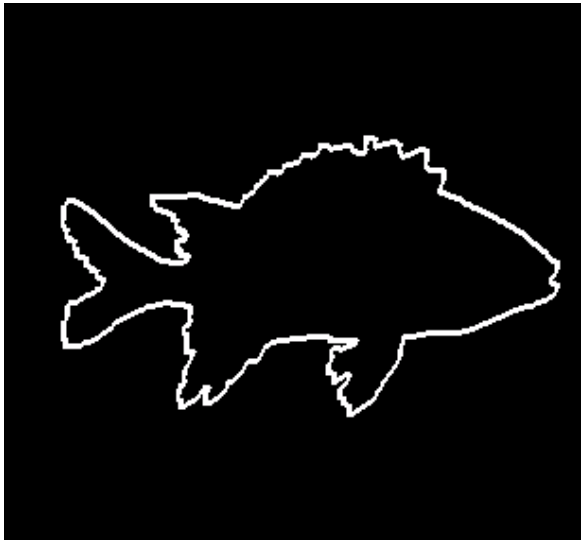
Angular Radial Transformation (ART)

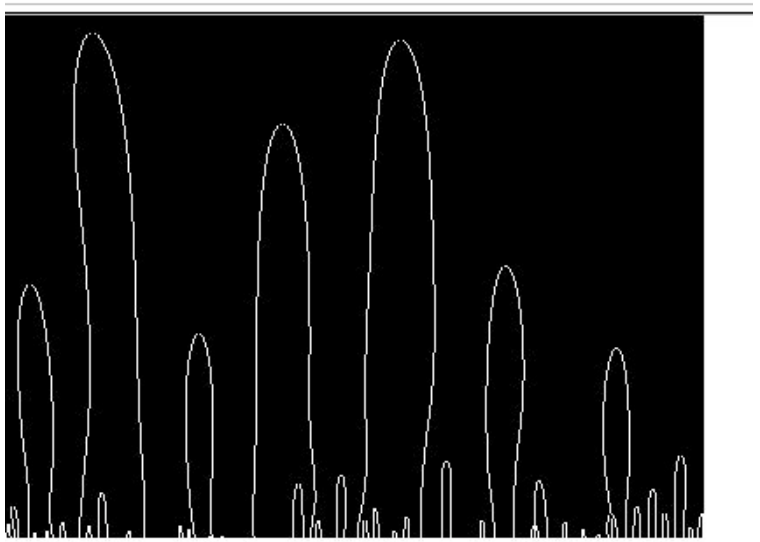
- Convert image information into angular and radial parts
- Represent image as coefficients of basis functions
- First 36 basis functions:



Curvature-Scale Space Computation

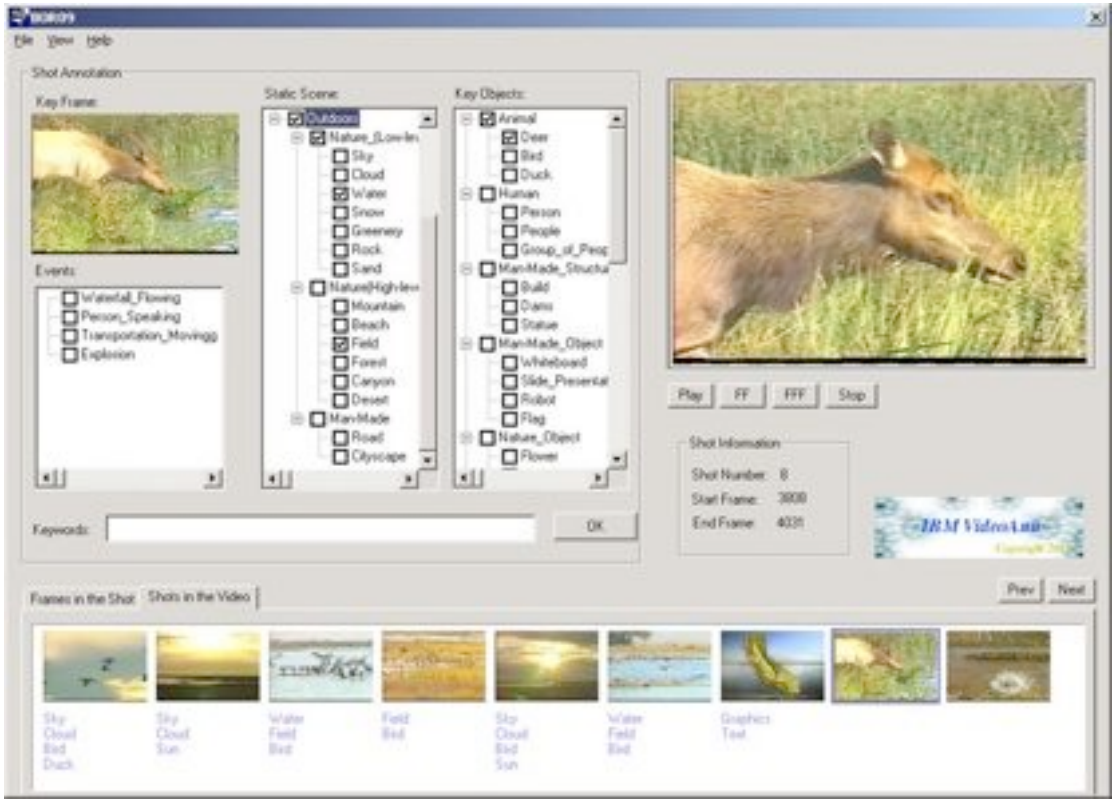
- Curvature is a local measure of how fast a curvature is turning
  - Curvature zero crossing points are essential for contours
  - Contour is sampled with increasing precision and smoothed stepwise to retrieve curvature zero-crossings of various scales
- Mokhtarian, Abbasi et al., University of Surrey, UK
  http://www.ee.surrey.ac.uk/CVSSP/demos/css/demo.html



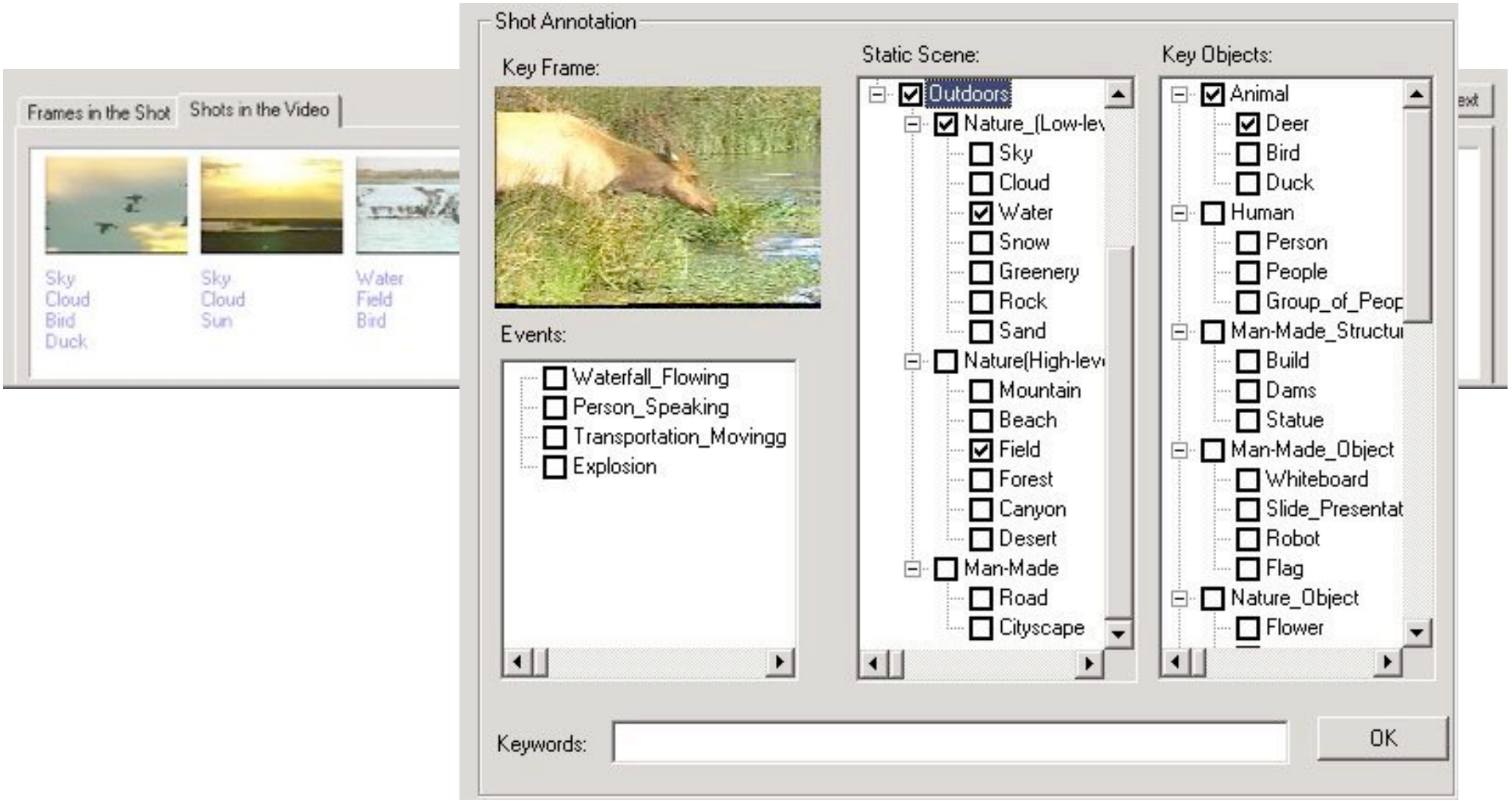


IBM VideoAnnEx (1)

- Support tool for manual annotation of video sequences with MPEG-7 metadata
  - Experimental tool 2001-2003, no longer supported
  - Requires a basic lexicon of description items in addition to video file



IBM VideoAnnEx (2)

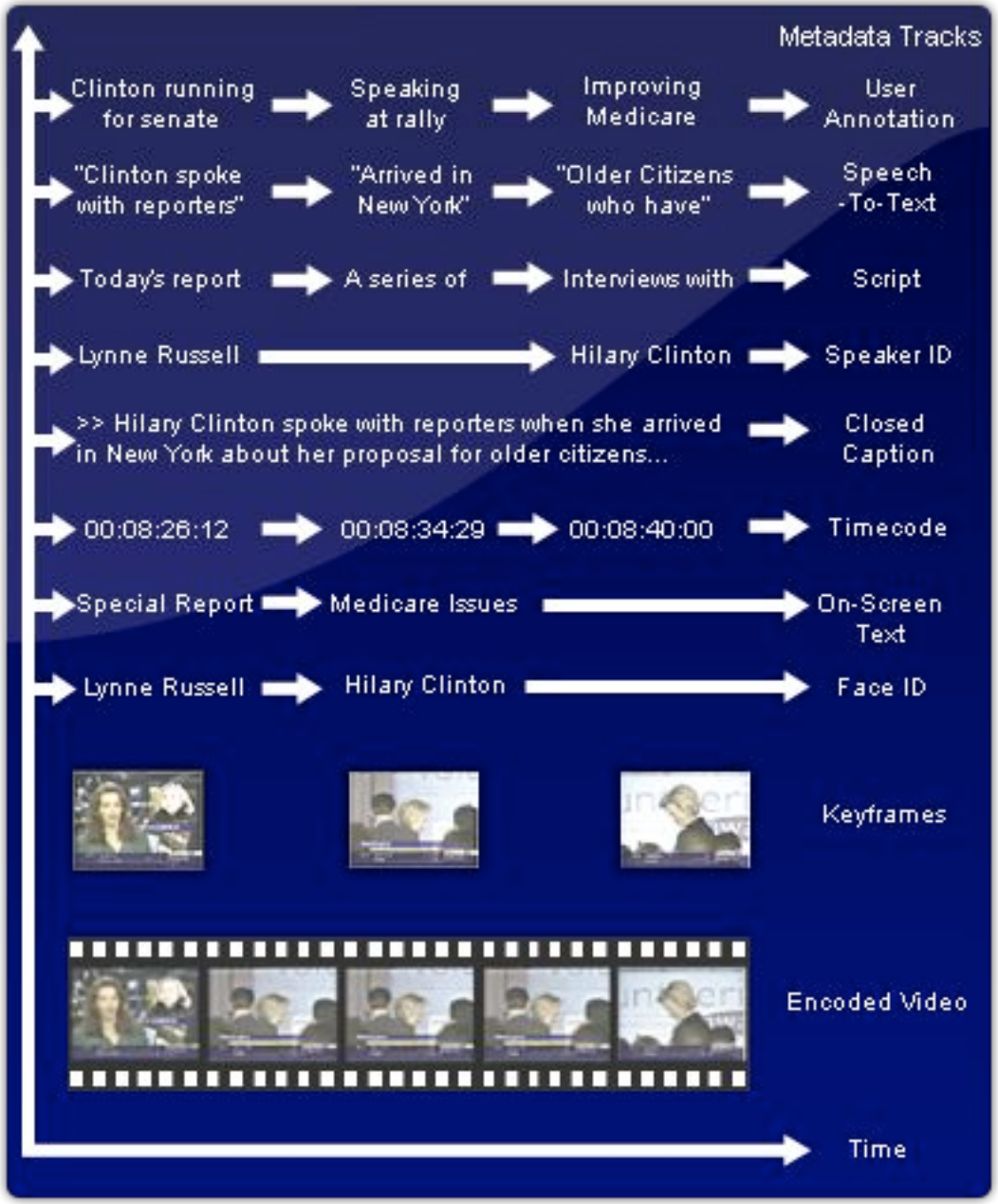


IBM VideoAnnEx (3)

Example: Autonomy Virage

www.virage.com:

"Using advanced image and audio analysis engines that watch, listen to and read a video signal in real-time, Autonomy Virage delves into the video file itself to extract the meaning of the information contained within."



Example: Autonomy Virage / DVI / IDOL

- DVI = Deep Video Indexing
- IDOL = Intelligent Data Operating Layer

  "Openly configurable, the DVI functionality uses numerous proprietary approaches in a configurable lattice which can be weighted within the DVI fingerprint as separate entities or as a whole unit.

  These technological approaches include:

  – Texture trajectory analysis
  – Advanced Optical Character Recognition (OCR)
  – Spectrum trajectory analysis
  – Advanced scene analysis

  Autonomy Virage can extract a comprehensive range of data from multimedia resources, including full transcripts of audio streams, on-screen character recognition, keyframes, facial recognition and speaker information, all of which is linked to the original video file, allowing users to locate content with pinpoint accuracy. Utilizing the power of IDOL, Autonomy Virage provides users with an unrivalled range of video analysis tools such as scene detection, 'find similar' functions and conceptual analysis such as automatic hyperlinking of related content, categorization and clustering."

  www.virage.com