8 Multimedia Content Description

Literature:

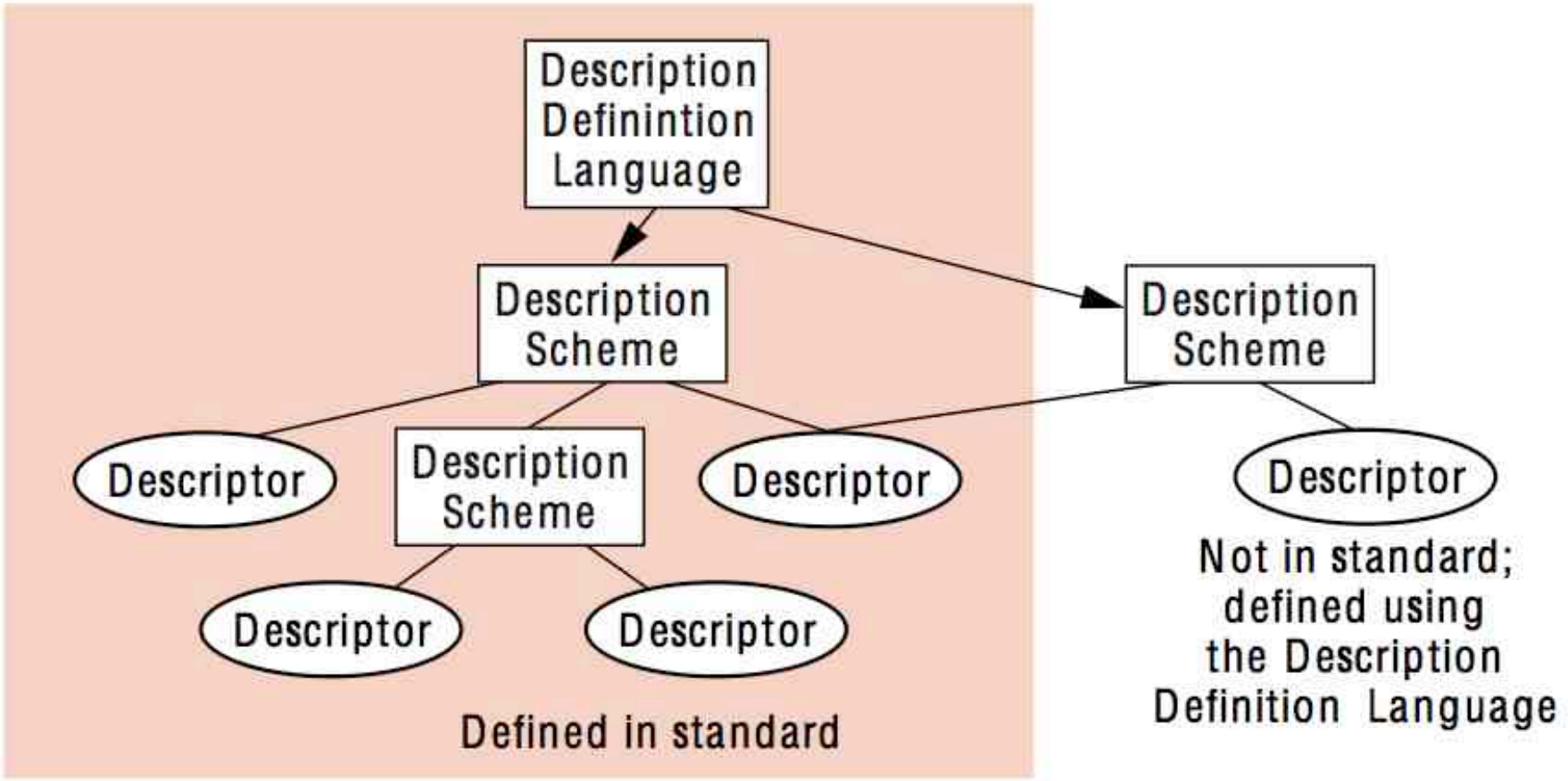
B.S. Manjunath et al. (eds.): Introduction to MPEG-7 - Multimedia Content Description Interface, Wiley 2002

MPEG-7 Overview, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

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

Metadata Classification in MPEG-7

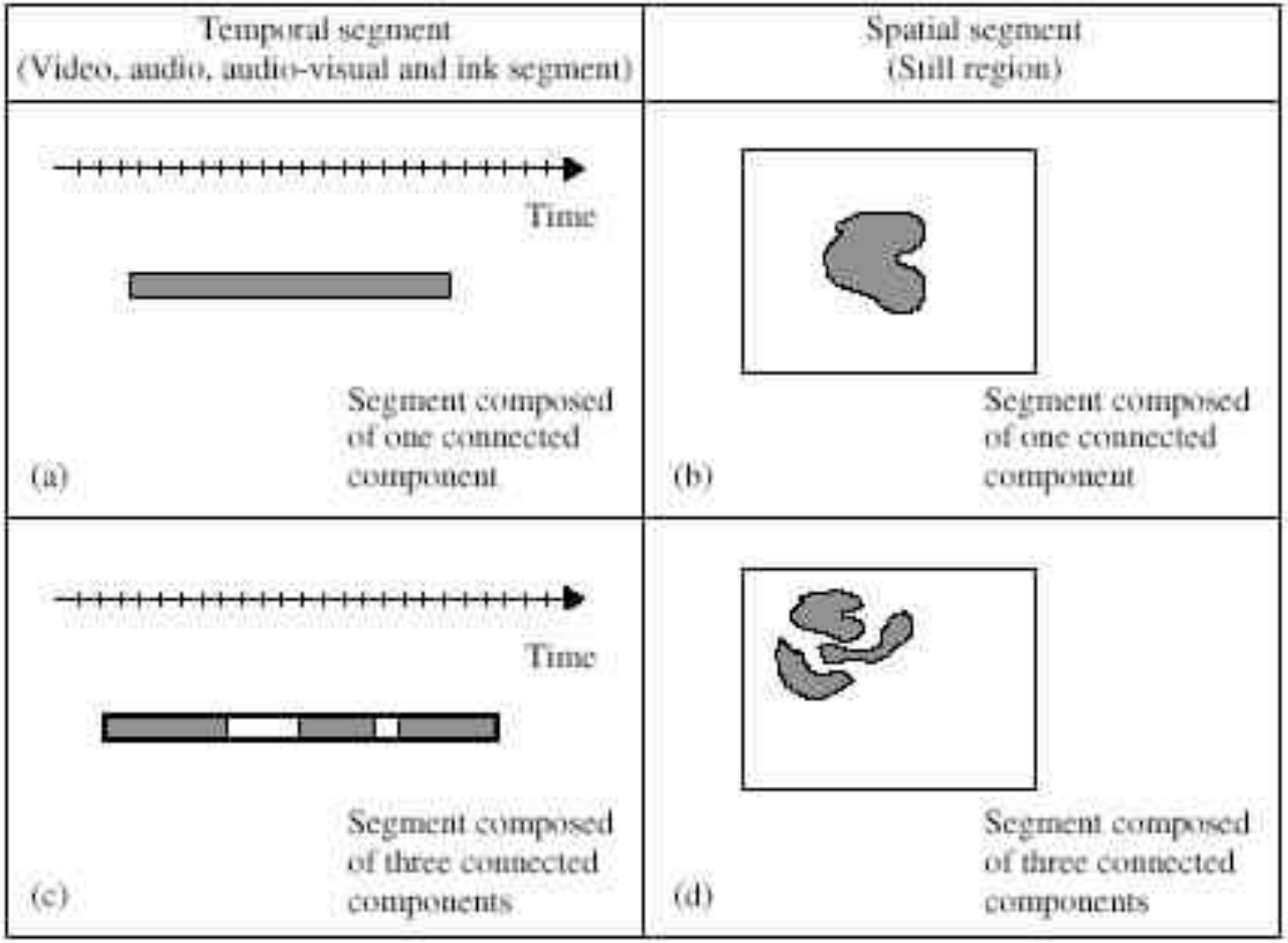
- Content Management
  - Media information (e.g. file name, format, resolution)
  - Creation information (e.g. creator, location, date)
  - Usage information (e.g. rights)
- Content Structure
  - Segments
  - Segment relations
- Content Semantics

Following slides: Details on Content Structure

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)
    - » ImageText DS
  - Video Segment DS (temporal)
    - » Analytic edited video segment DSs
  - Moving Region DS (spatiotemporal)
    - » VideoText DS
  - 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

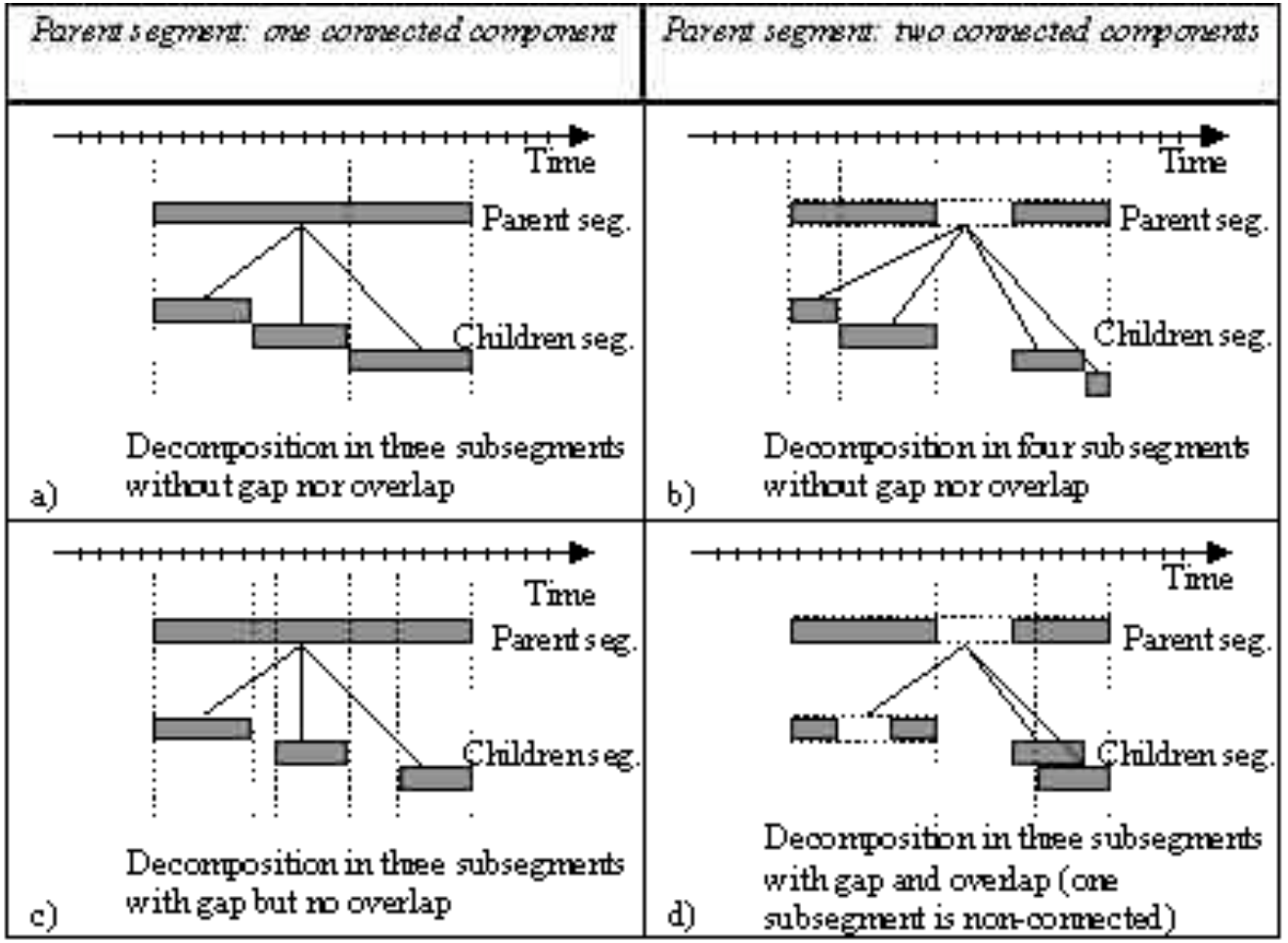


Segment Attributes

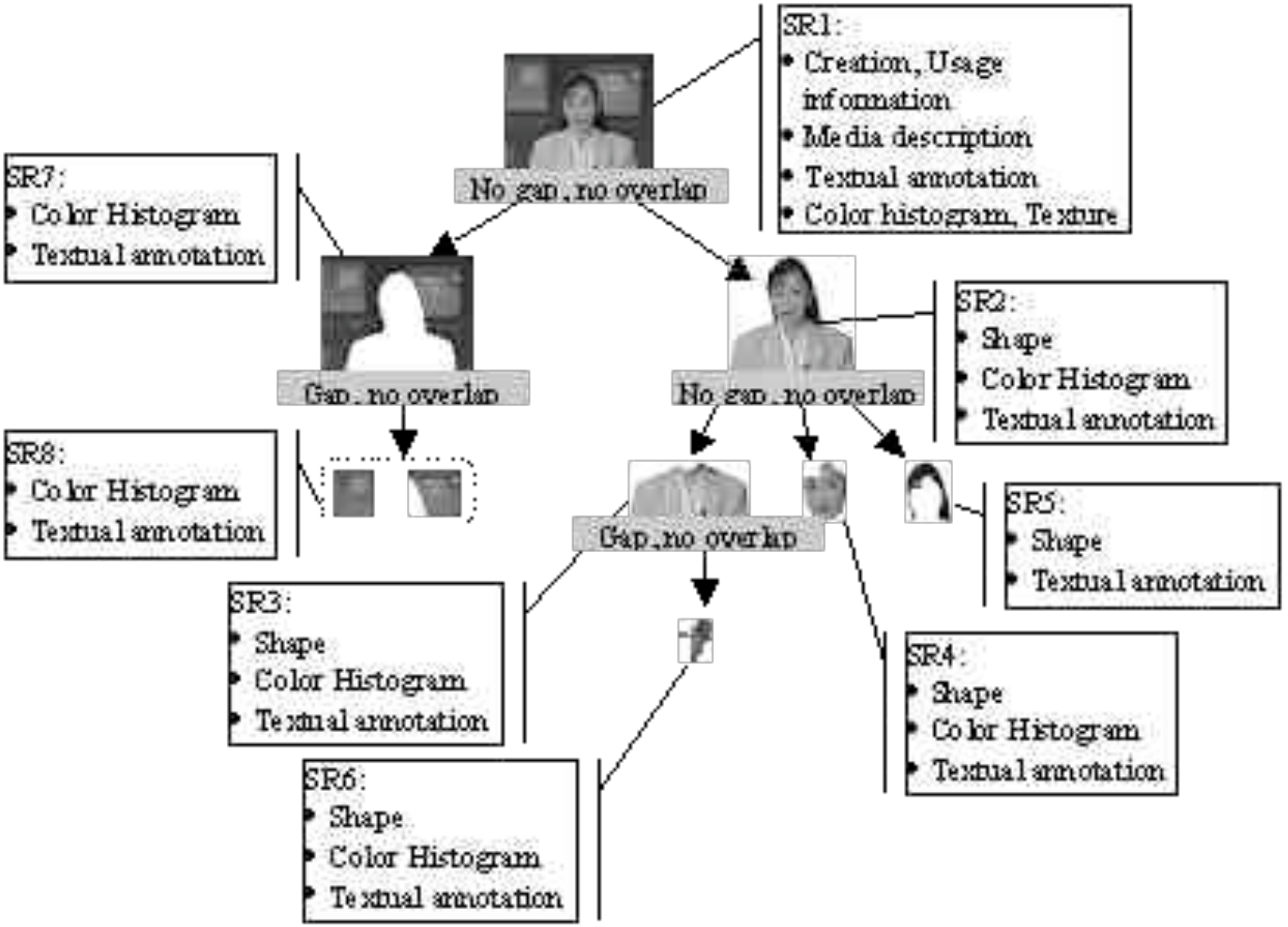
- Generic features
  - (media information, creation information, usage information, annotations)
- Media type dependent features:
  - (visual features, audio features)
- Specific features for segments
  - Mask Descriptor
    - » Spatial mask, Temporal mask, Spatio-temporal mask
  - Importance of descriptors
    - » MatchingHint: relative importance of descriptors
    - » PointOfView: relative importance of segments for a specific point of view (PointOfView given as string, e.g. “Home team” for soccer game)
  - Ink segment descriptors
    - » Handwriting recognition information (recognizer, lexicon)
    - » Handwriting recognition result (quality, accuracy-scored results)

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



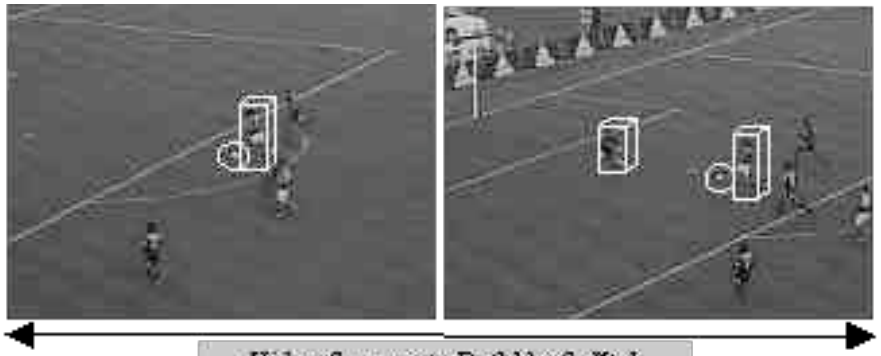
Example of Image Description



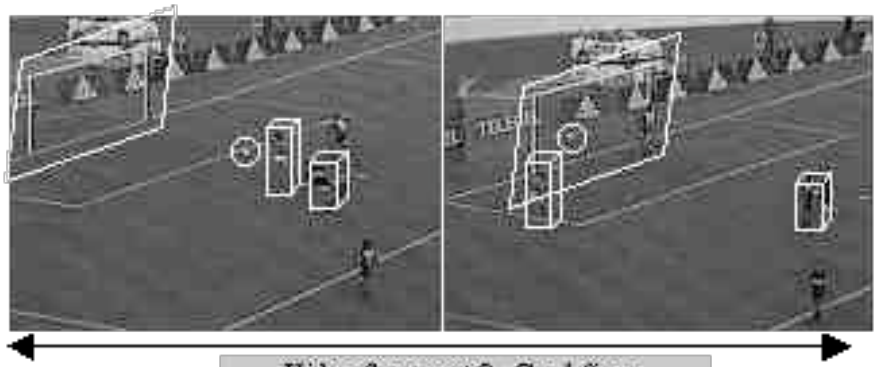
Structural Relations of Segments

- 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.

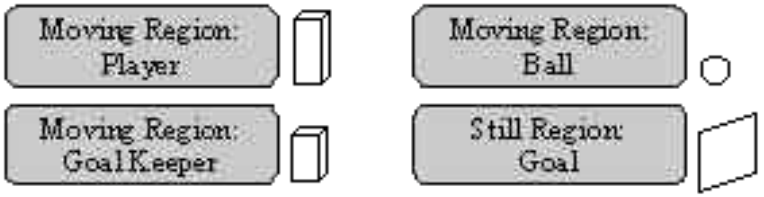
Video Segmentation with Moving Regions

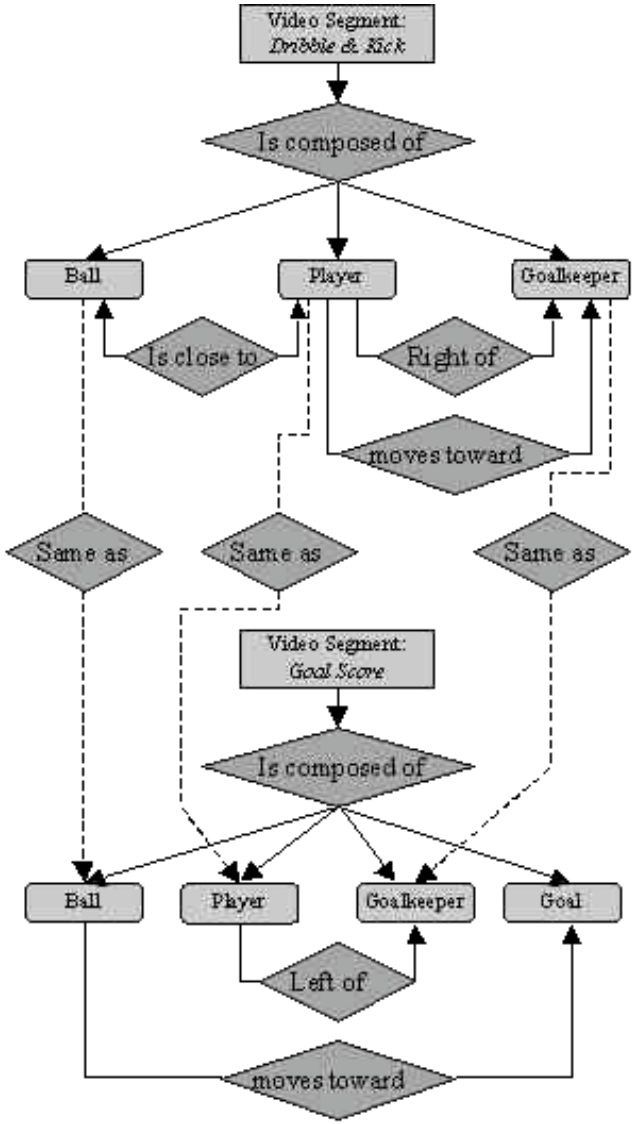


Video Segment: *Dribble & Kick*



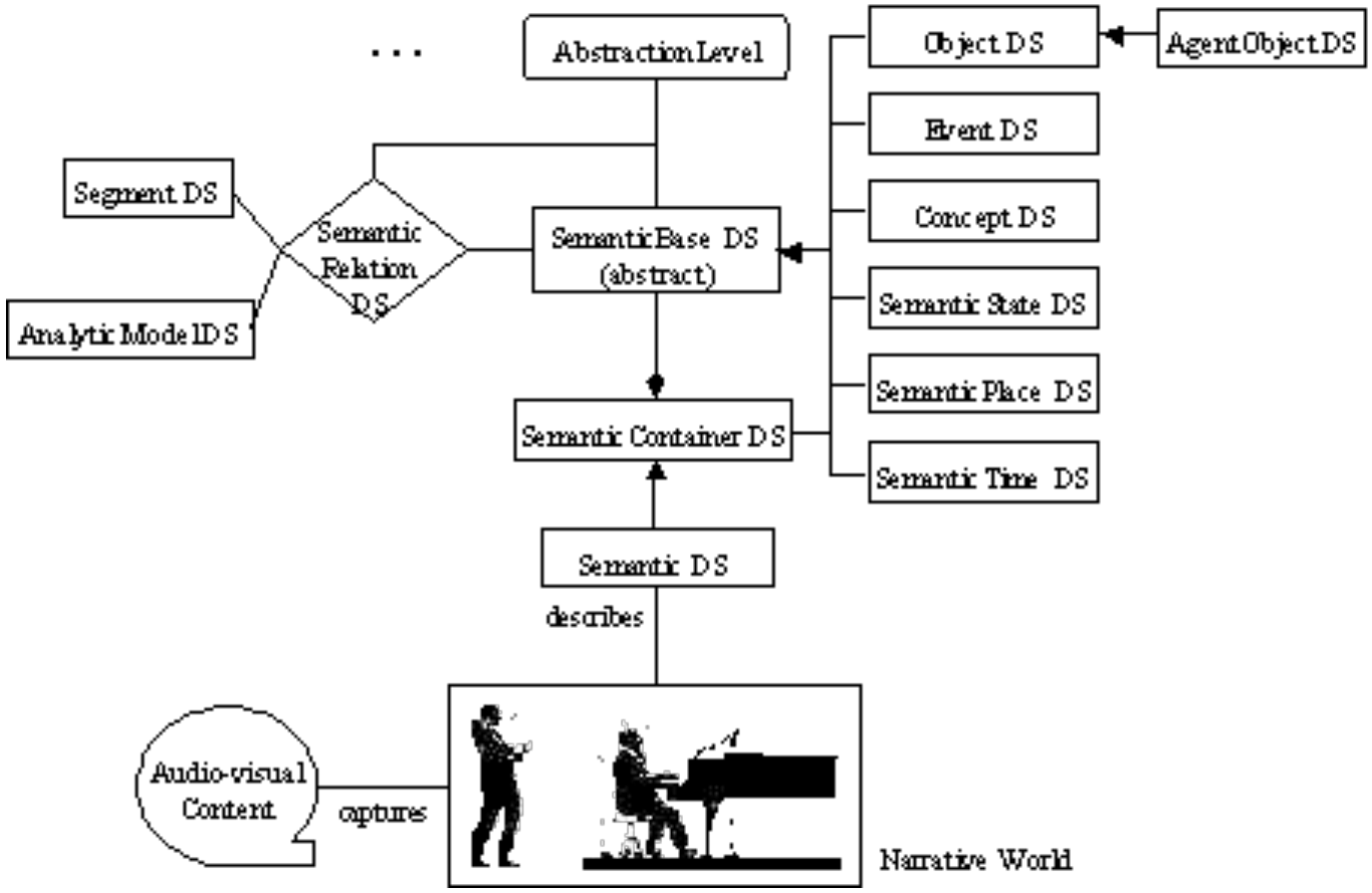
Video Segment 2: *Goal Score*



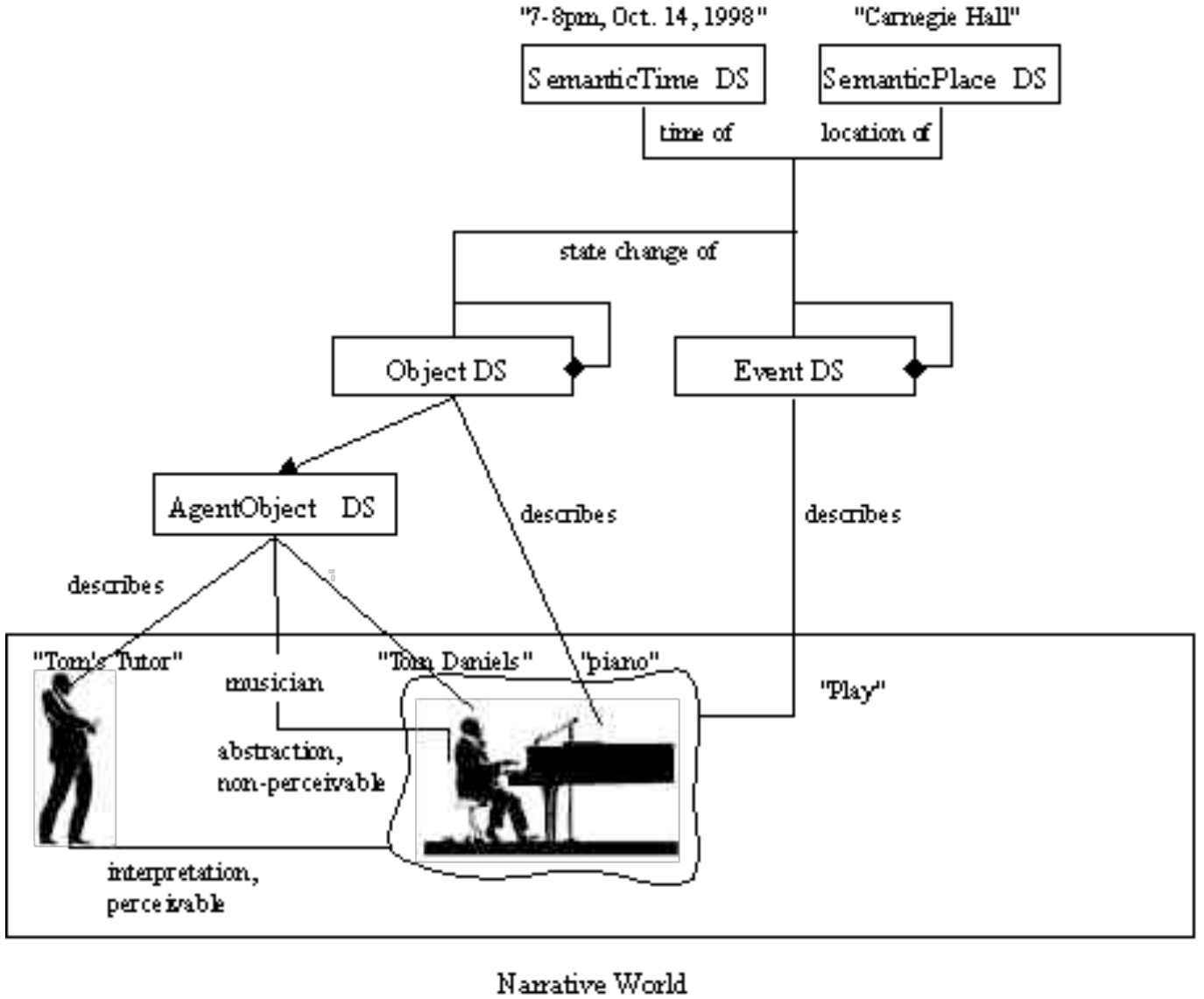


Content Semantics in MPEG-7

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

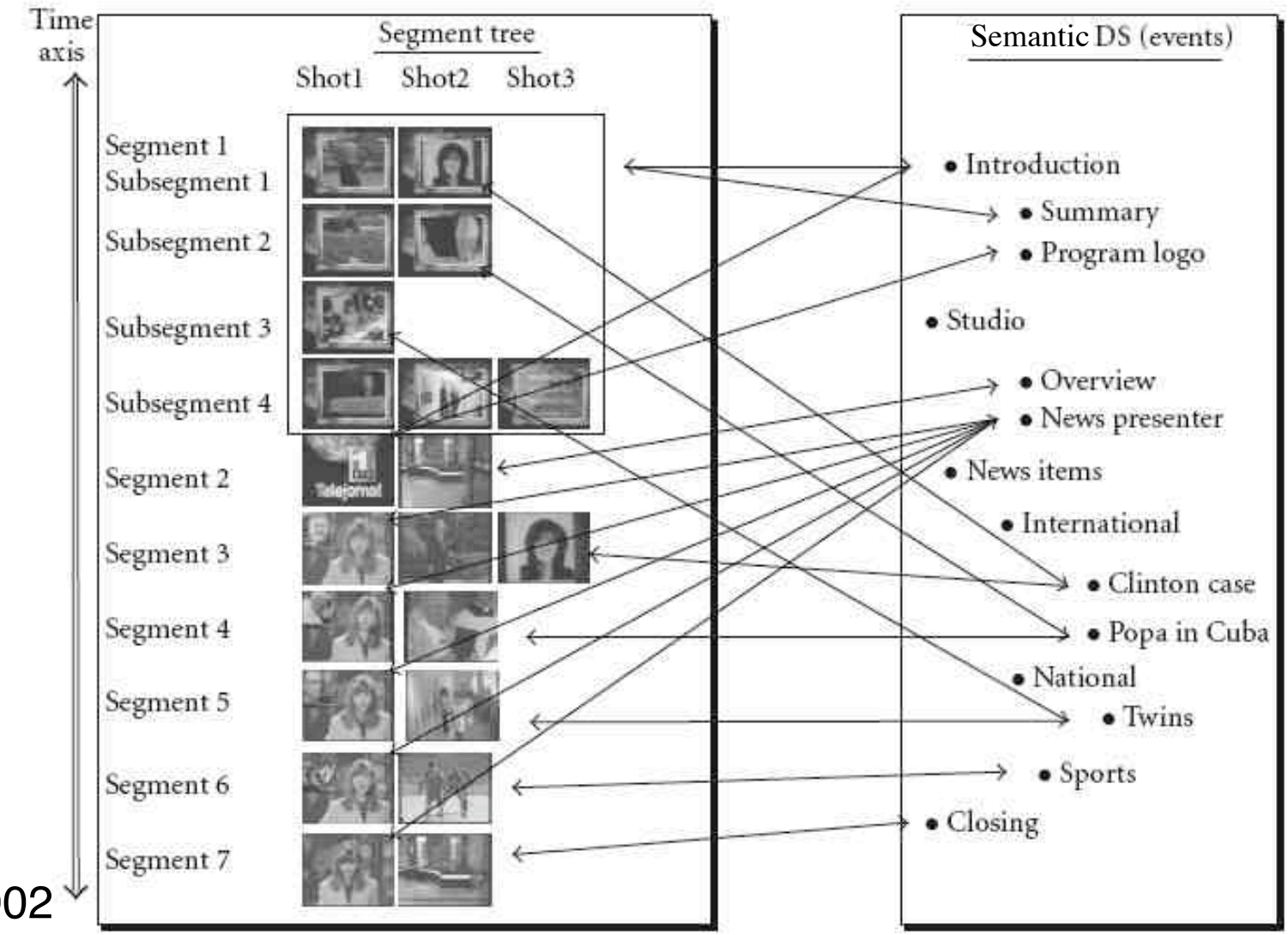


Content Semantics in MPEG-7: Example



Relating Structure and Semantics: Example

- Creating a semantic index for a video sequence

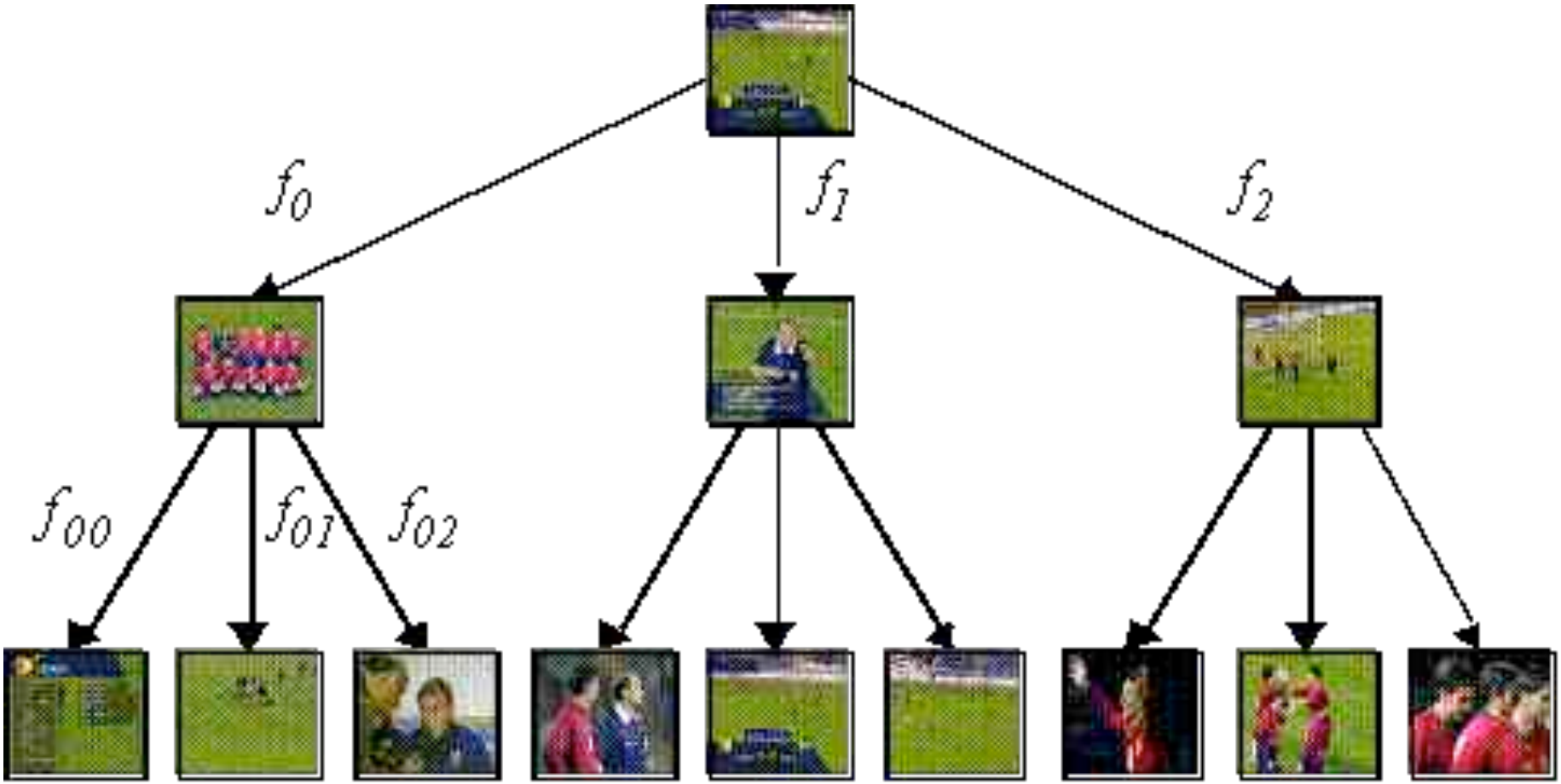


Salembier 2002

Navigation and Access

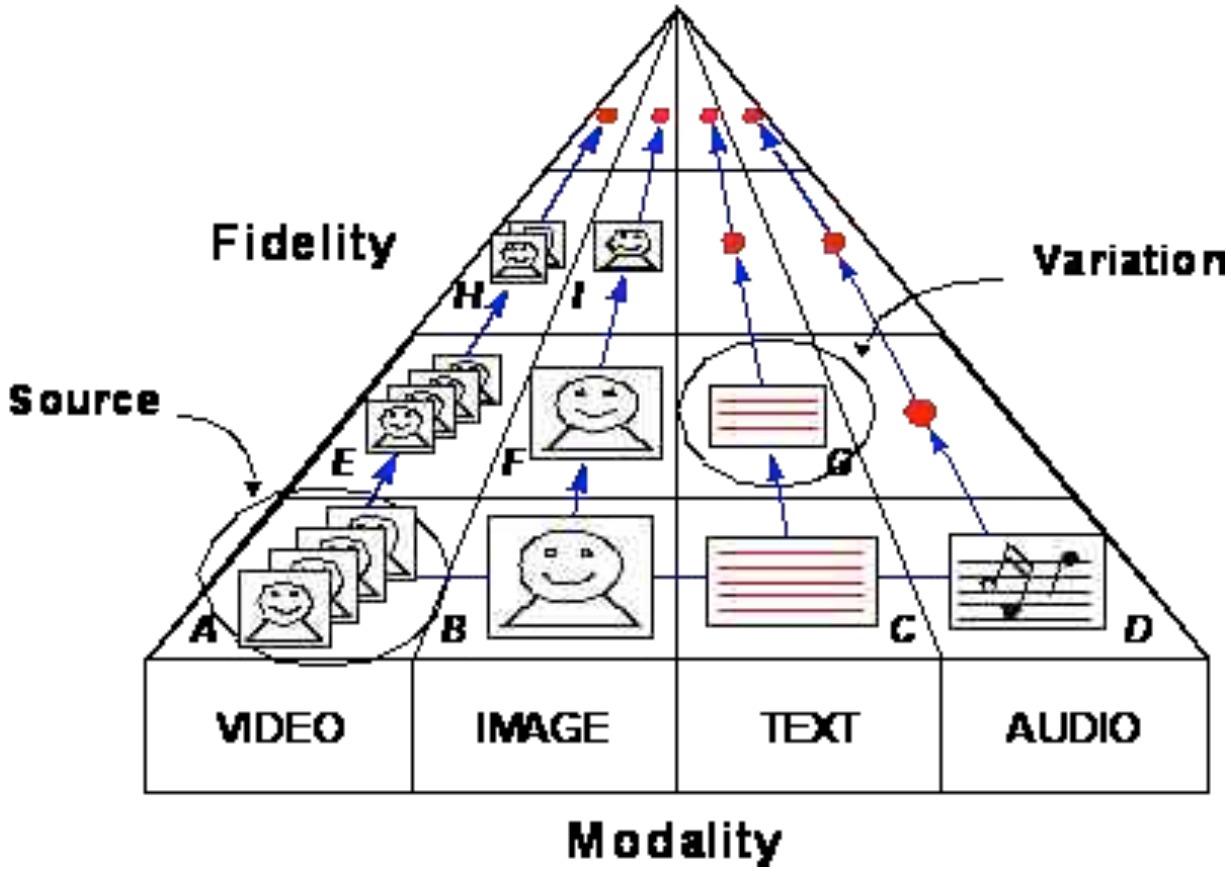
- Description schemes to facilitate navigation and access of audio-visual content:
  - Summaries
    - » Browsing, navigation, discovery, visualization, sonification
  - Views and partitions
    - » Representations in multiple domains, resolutions
  - Variations
    - » Different versions adapted to delivery conditions

Example: Summary as Hierarchy of Key Frames



Variations

- Components of a complex multimedia object may exist in various variations (different resolutions, languages, etc.)
  - Server or proxy server should be able to select the appropriate variation



MPEG-7 Visual Description Tools

- Descriptors for the following basic visual features:
  - Color, Texture, Shape, Motion, Localization, and Face recognition
  - Each category consists of elementary and sophisticated Descriptors
- Basic structures for composing visual features:
  - Grid layout
  - Time series
  - Multiple (2D/3D) view
  - Spatial 2D coordinates
  - Temporal interpolation

Principles of Automatic Feature Extraction

- Colour:
  - Histogram, colour clusters
- Texture:
  - Spectral distribution, energy
- Motion:
  - Vector histogram, parametric models
- Contours:
  - Moments, wavelet coefficients
- Faces:
  - Vector basis and similarity matching

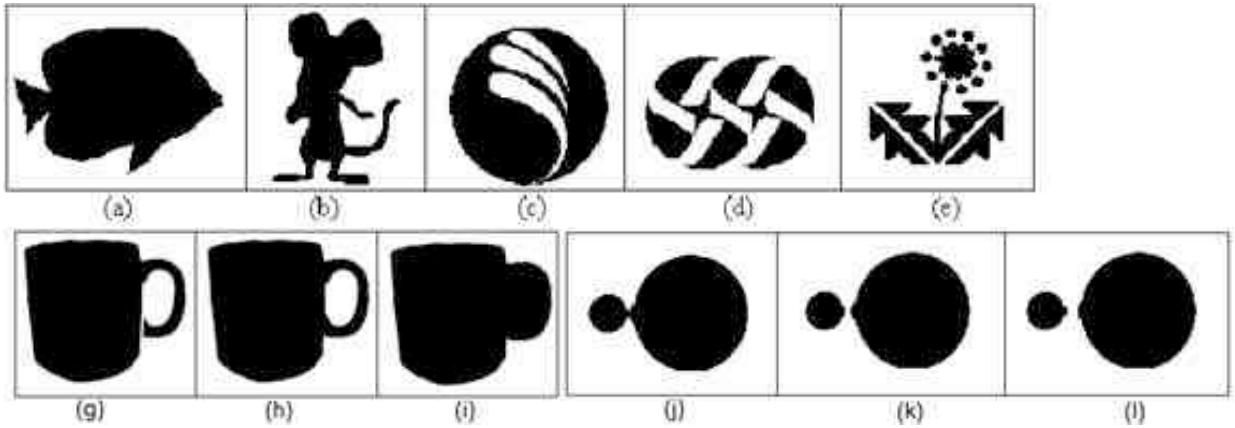
- Usage of compressed data formats:
  - E.g. frequency space transformation (JPEG), motion estimation (MPEG-2) can be re-used for feature extraction

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

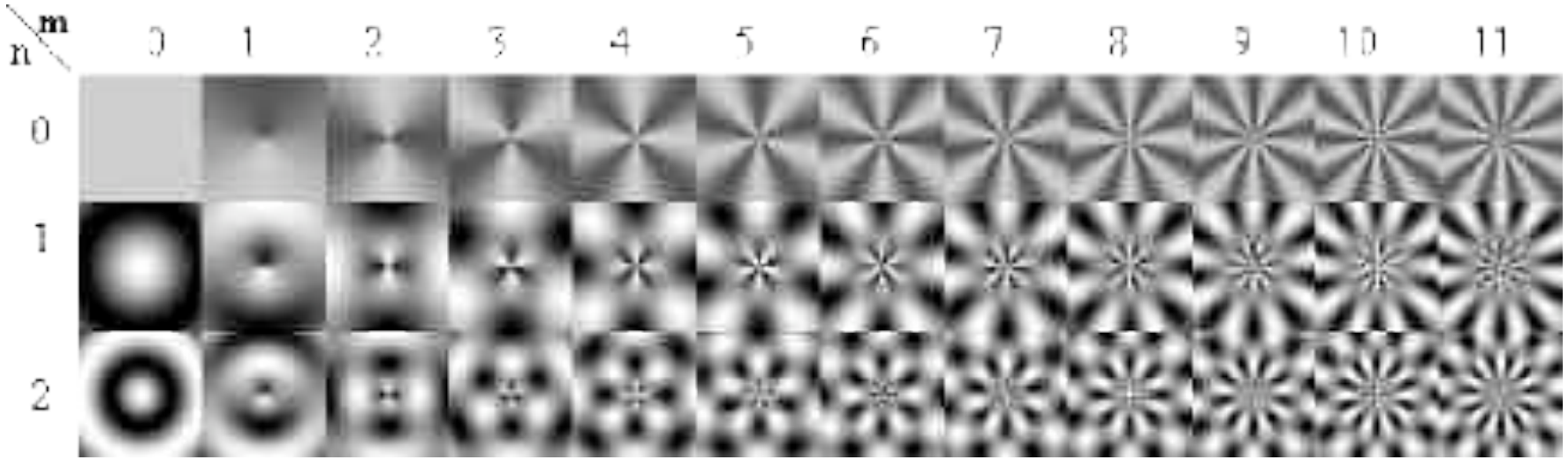
Region shapes:



Contour shapes:

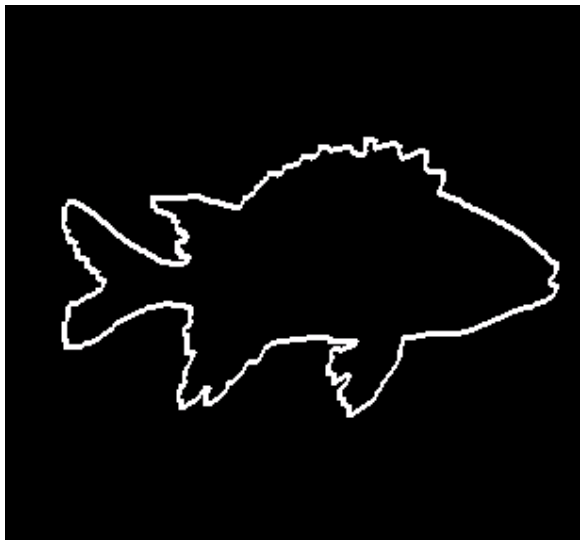
Angular Radial Transformation (ART)

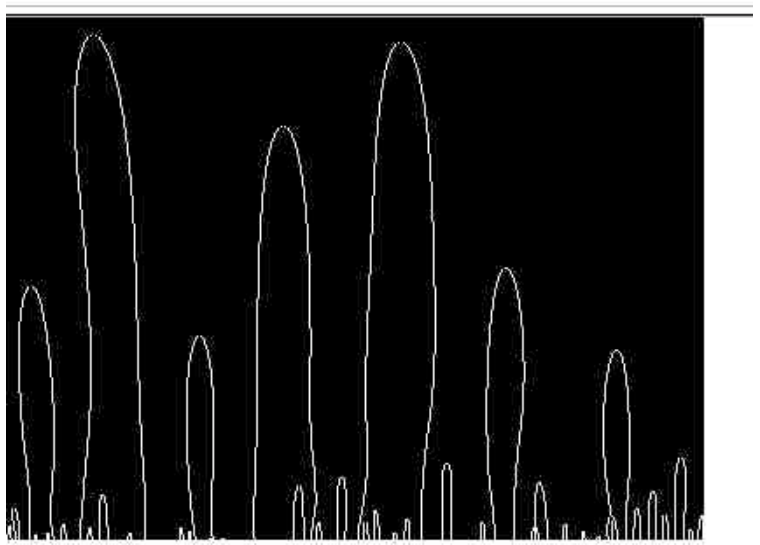
- Convert image information into angular and radial parts
- Represent image as coefficients of basis functions
- First 36 basis functions:



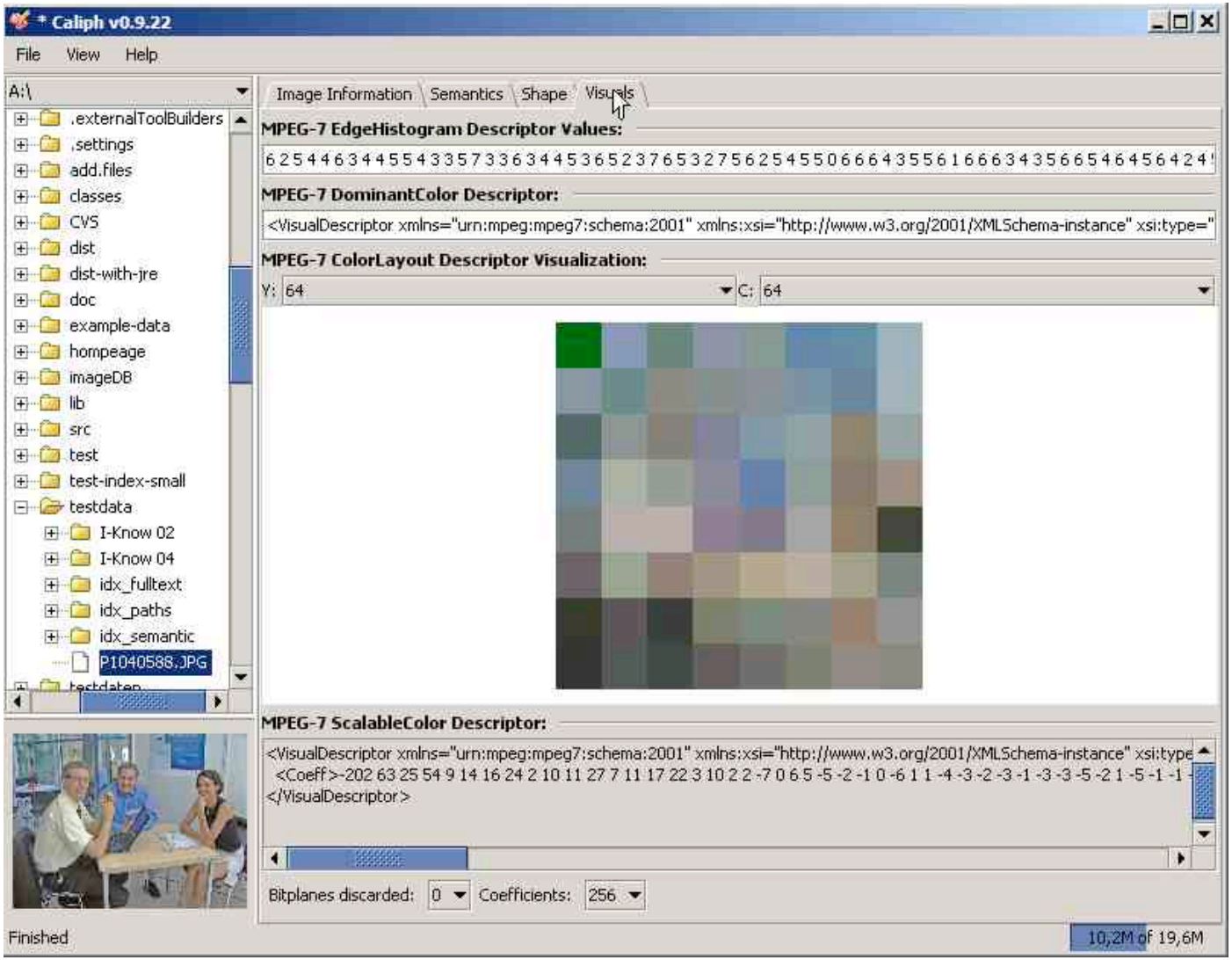
Curvature-Scale Space Computation

- 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



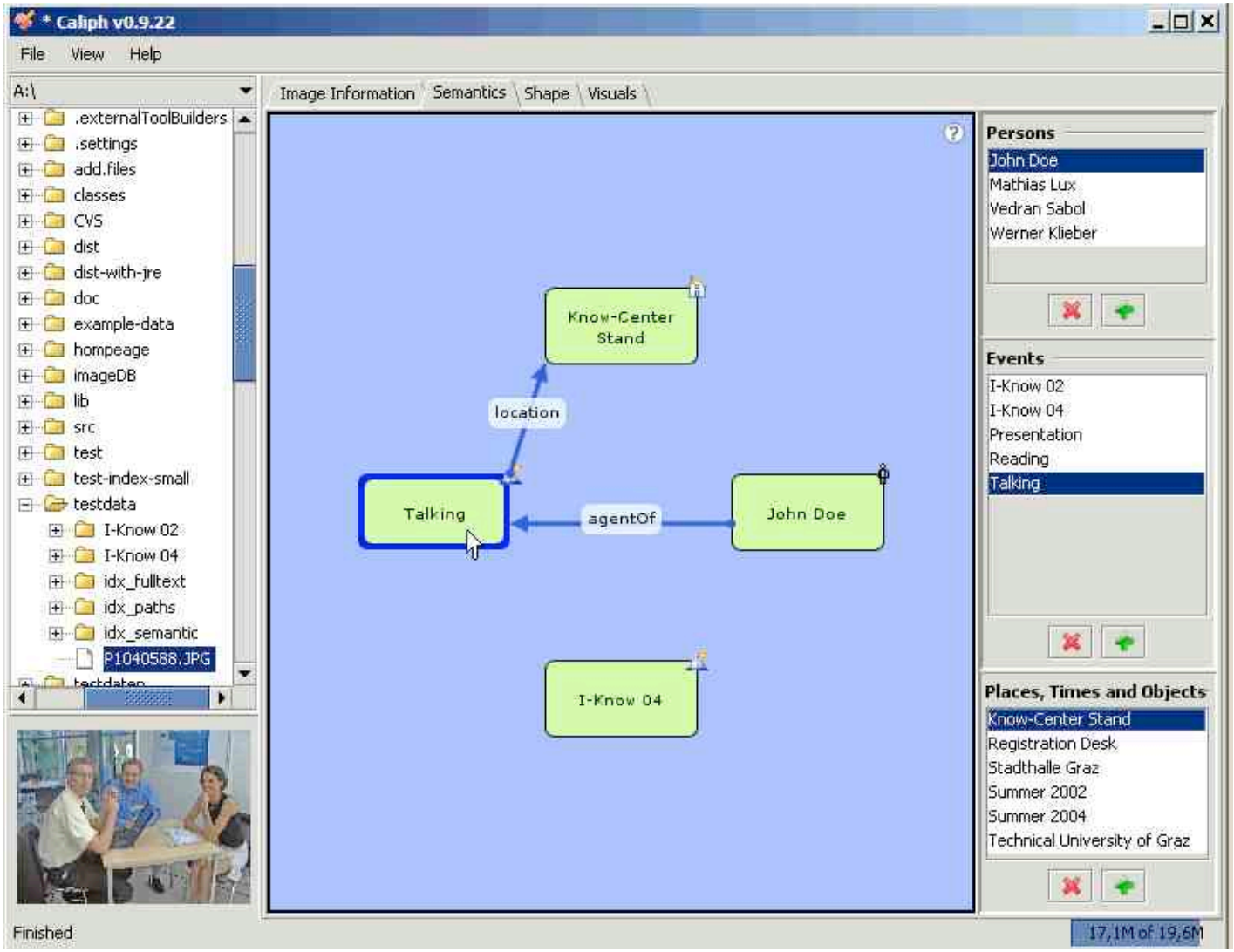


Example: Still Image Annotation Tool Caliph (1)



Mathias Lux,
Caliph & Emir
(*ACM Multimedia* 2009)

Example: Still Image Annotation Tool Caliph (2)

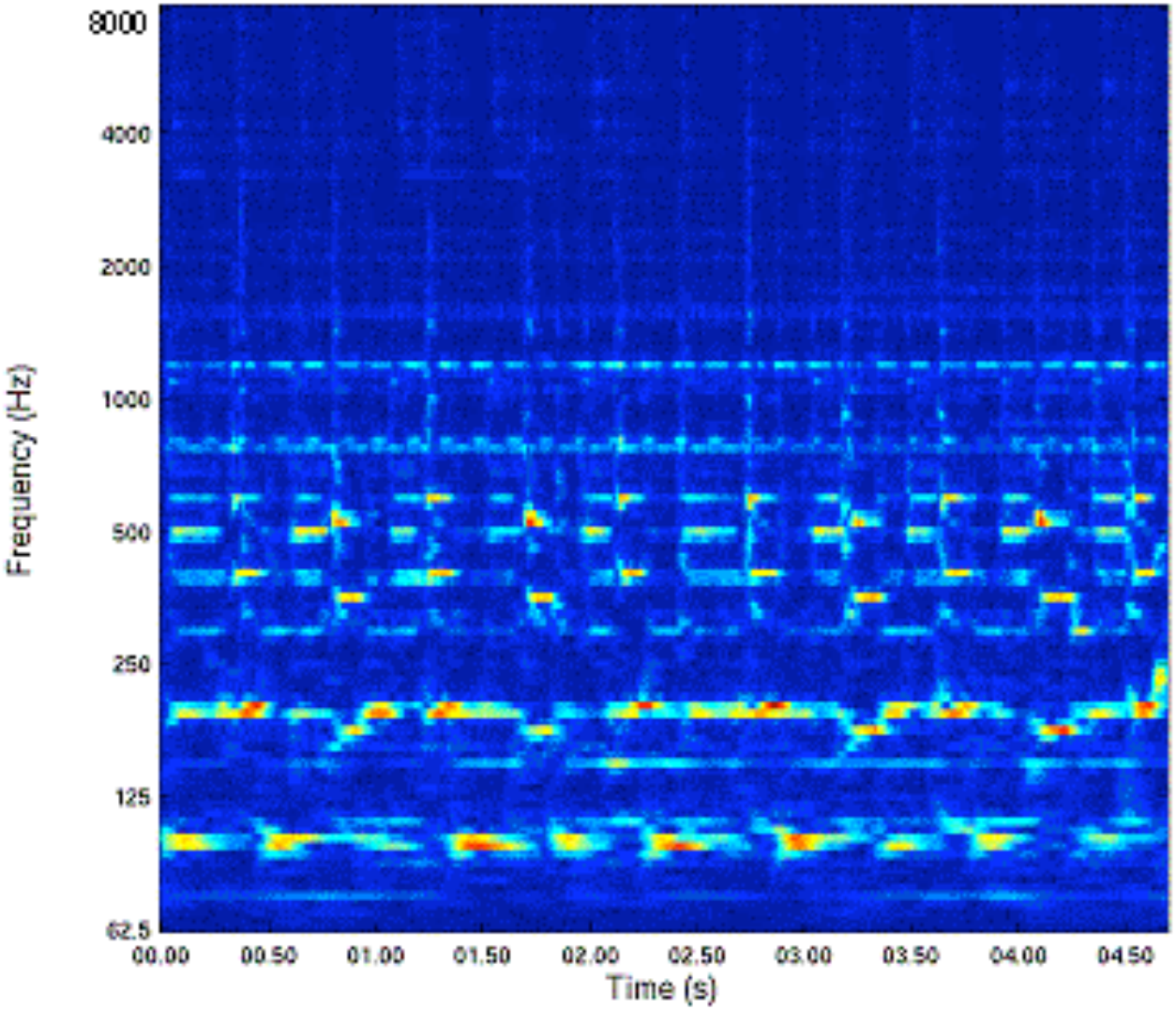


www.semanticmetadata.net

MPEG-7 Audio Description Tools

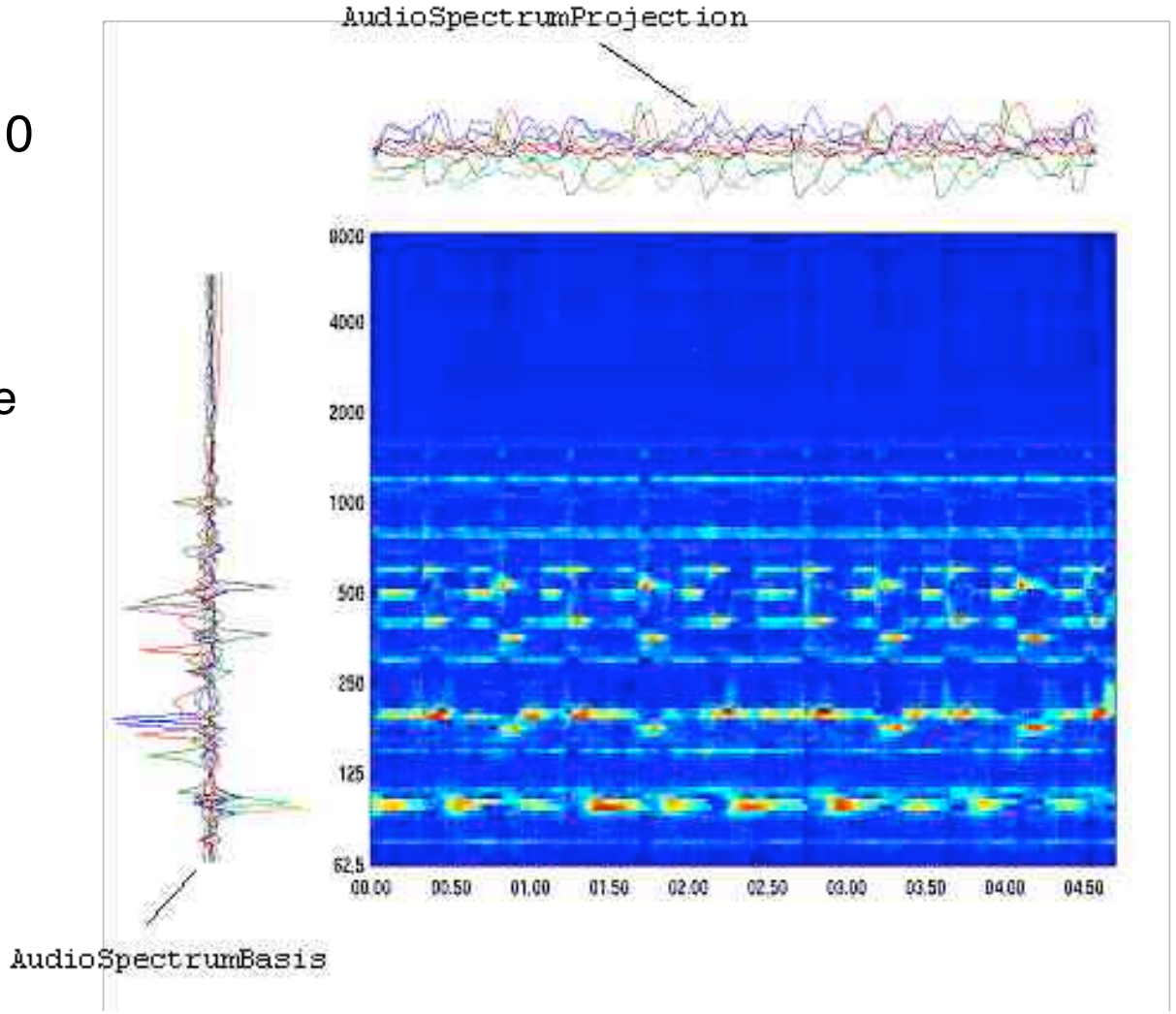
- Low-level audio descriptors:
  - Basic: Instantaneous waveform and power values
  - Basic spectral: Log-frequency power spectrum and spectral features (centroid, spread, flatness)
    - » AudioSpectrumEnvelope: Spectrogram of the signal
  - Signal parameters: Fundamental frequency
  - Temporal timbral: Log attack time and temporal centroid
  - Spectral timbral: Specialized spectral features
- High-level audio descriptors:
  - Sound recognition and indexing
  - Musical instrument timbre description
  - Melody description tools
  - Spoken language recognition

Spectral Analysis with AudioSpectrumEnvelope



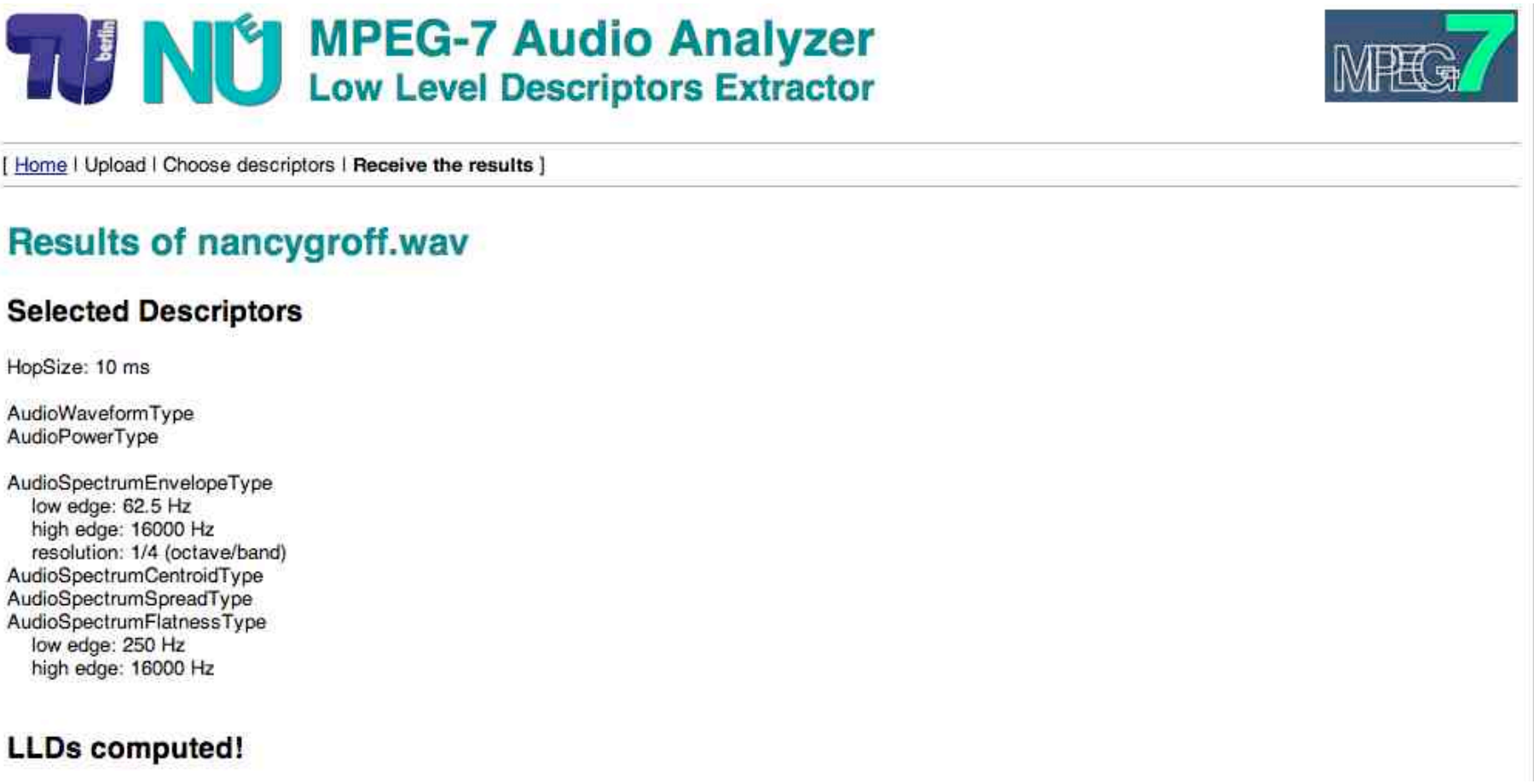
Data-Reduced Spectral Representation

- Reconstruction of sonogram using a compact representation of 10 vectors
  - required storage space 10(M+N) values
  - M number of time points
  - N number of spectrum bins



Example: Automated Audio Feature Extraction (1)

http://mpeg7lld.nue.tu-berlin.de/



Example: Automated Audio Feature Extraction (2)

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!-- TU Berlin Audio Analyzer v1.0:
http://www.nue.tu-berlin.de/forschung/projekte/mpeg7/ -->
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
    xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
<Description xsi:type="ContentEntityType">
<MultimediaContent xsi:type="AudioType">
<Audio xsi:type="AudioSegmentType">

<AudioDescriptor xsi:type="AudioWaveformType"  minRange="-0.734375"
maxRange="0.671875">
    <SeriesOfScalar hopSize="PT10N1000F" totalNumOfSamples="1589" >
        <Min> 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
        0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
        0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
        0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
        0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
        0.000000 0.000000 -0.070313 -0.179688 -0.250000 -0.257813 -0.210938
        -0.218750 -0.226563 -0.234375 -0.210938 -0.210938 -0.218750 -0.203125
        -0.195313 -0.164063 -0.164063 -0.164063 -0.148438 -0.132813 -0.117188
        -0.125000 -0.109375 -0.109375 -0.109375 -0.101563 -0.109375 -0.101563
        -0.101563 -0.203125 -0.171875 -0.148438 -0.117188 -0.140625 -0.132813
        -0.156250 -0.140625 -0.203125 -0.117188 -0.164063 -0.117188 -0.156250
        -0.156250 -0.187500 -0.187500 -0.132813 -0.132813 -0.156250 -0.156250
        -0.132813 -0.132813 -0.148438 -0.148438 -0.203125 -0.203125 -0.156250
        -0.156250 -0.148438 -0.156250 -0.140625 -0.171875 -0.164063 -0.179688
        -0.132813 -0.164063 -0.125000 -0.164063 -0.125000 -0.171875 -0.125000
        -0.156250 -0.101563 -0.156250 -0.289063 -0.289063 -0.500000 -0.500000
        -0.468750 -0.468750 -0.359375 -0.359375 -0.281250 -0.281250 -0.304688
        -0.304688 -0.304688 -0.304688 -0.226563 -0.226563 -0.218750 -0.218750
```

Example: Automated Audio Feature Extraction (3)

MPEG-7
Spoken Content Demonstrator

SpokenContentTranscription:

Index	Phone Label	Start Time (x10ms.)	End Time (x10ms.)
0	...	0	1
1	h	1	5
2	OY	5	16
3	k	16	26
4	t	26	32
5	E	32	40
6	l	40	42
7	I	42	51
8	s	51	63
9	f	63	67
10	Q	67	74
11	aI	74	
12	n	83	
13	S	88	
14	2:	102	
15	n	117	
16	6	122	
17	k	130	
18	...	134	
19	t	139	
20	h	144	
21	a:	150	
22	k	164	
23	...	168	
24	k	175	
25	...	183	
26	...	189	

```
<Block num="0" audio="unknown" defaultSpeakerInfoRef="#SpeakerX">
<MediaTime>
<MediaTimePoint>2003-11-10T00:00:00</MediaTimePoint></MediaTime>
<Node num="0" timeOffset="0" >
<PhoneLink nodeOffset="1" probability="1.000000e+000" acousticScore="-6.290000e+001" phone="0"/>
</Node>
<Node num="1" timeOffset="1" >
<PhoneLink nodeOffset="1" probability="2.497200e-002" acousticScore="-2.908300e+002" phone="16"/>
</Node>
<Node num="2" timeOffset="5" >
<PhoneLink nodeOffset="1" probability="9.466462e-003" acousticScore="-7.339000e+002" phone="39"/>
</Node>
<Node num="3" timeOffset="16" >
<PhoneLink nodeOffset="1" probability="9.536916e-002" acousticScore="-7.209400e+002" phone="6"/>
</Node>
```

8 Multimedia Content Description

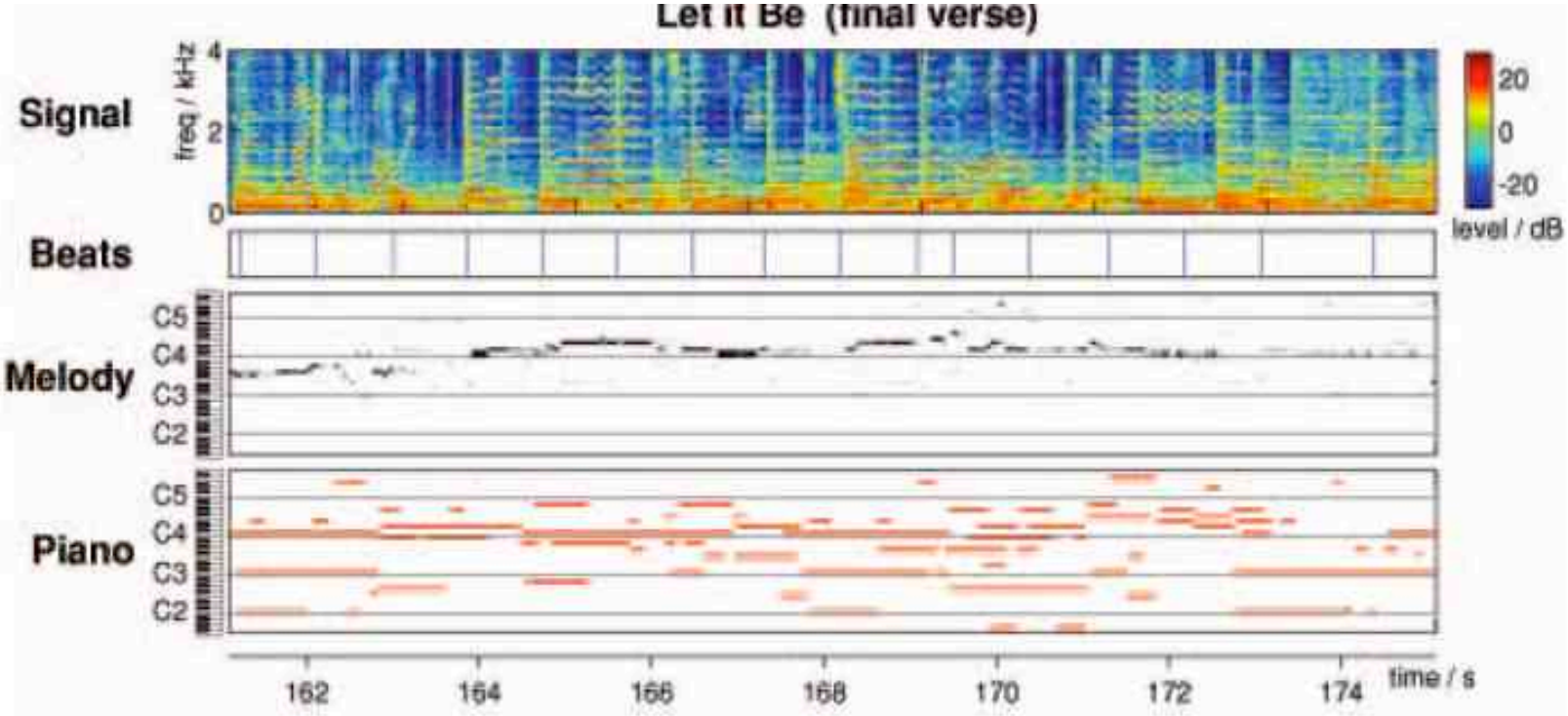
Literature:

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”)

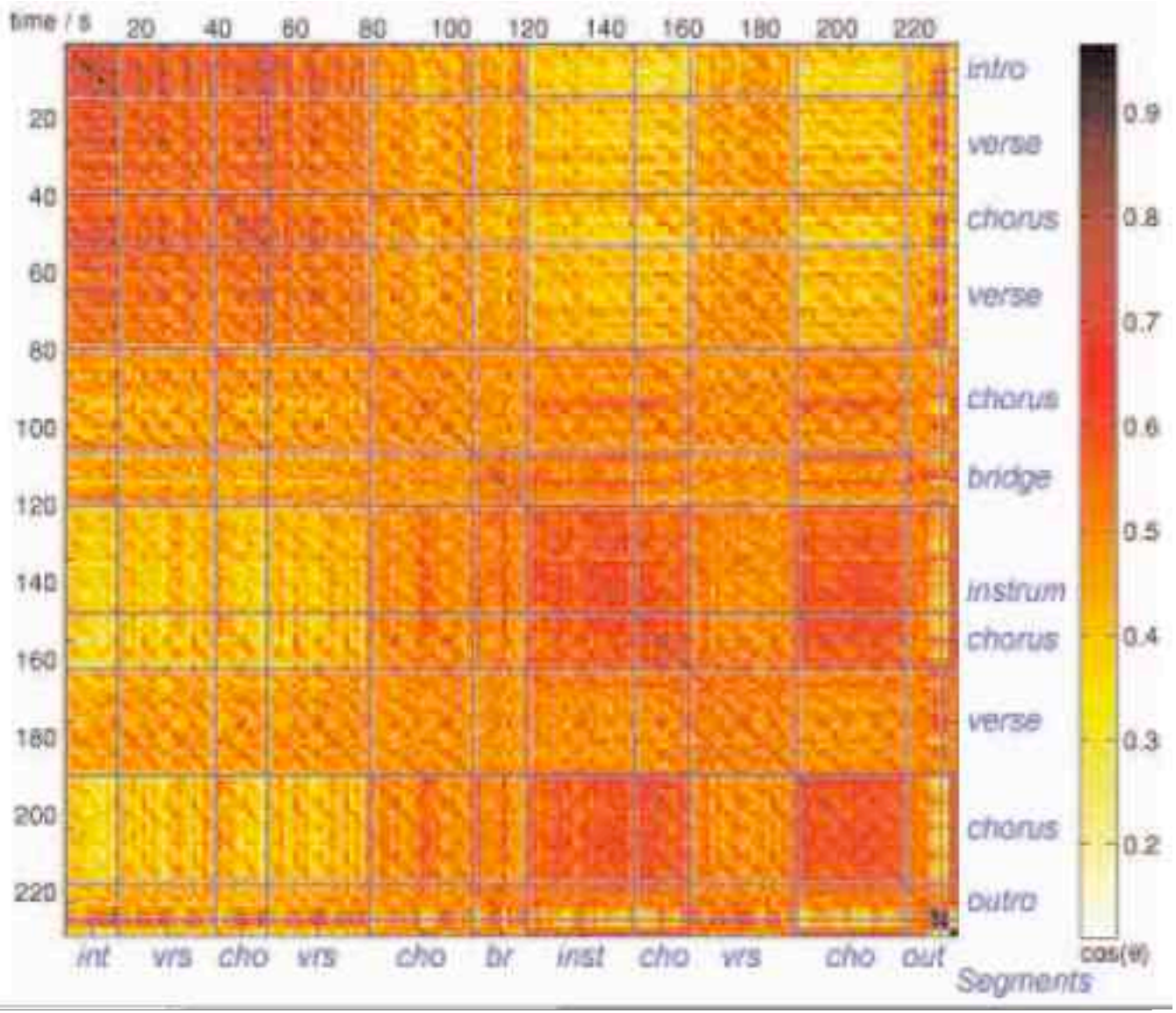
Automatic Score Transcription



- 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
  - Looking for diagonal ridges off the main diagonal
  - Blue lines are manually inserted for comparison



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)

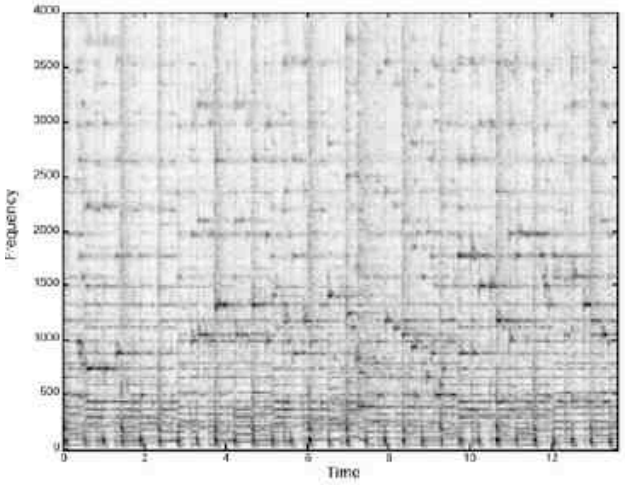


Fig. 1A - Spectrogram

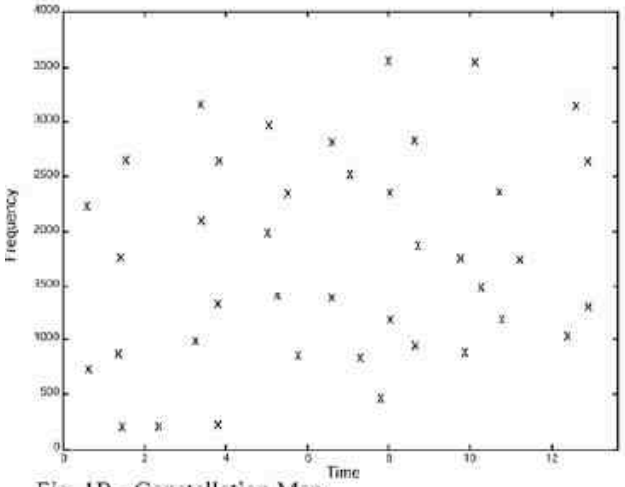


Fig. 1B - Constellation Map

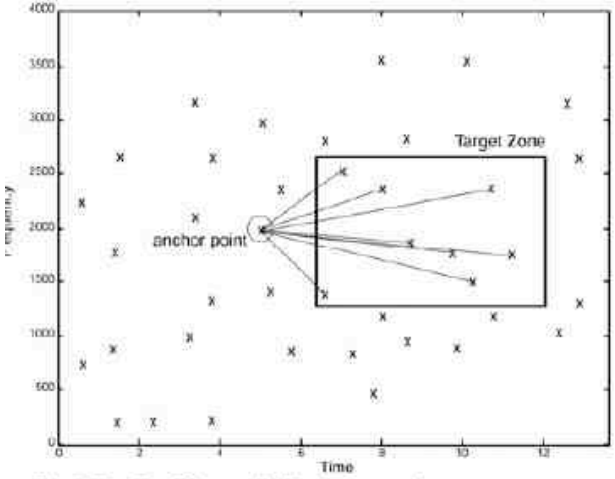


Fig. 1C - Combinatorial Hash Generation

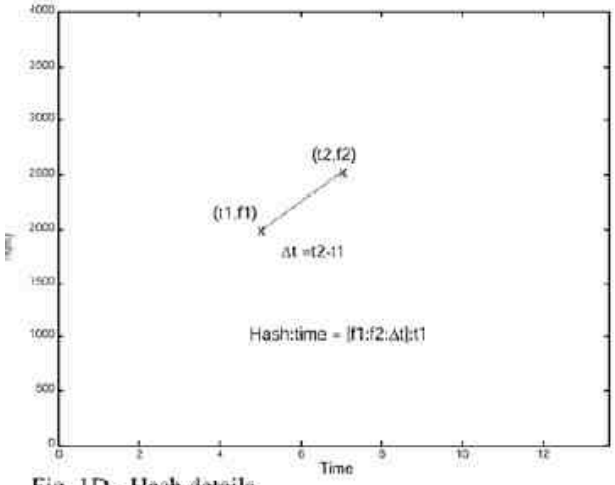


Fig. 1D - Hash details

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

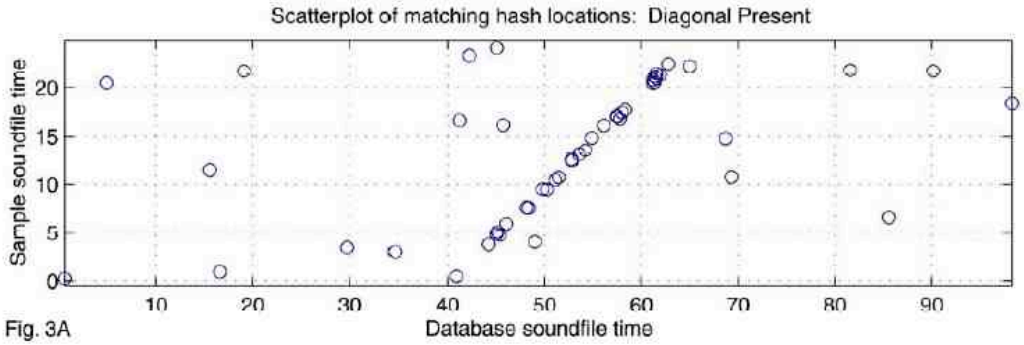


Fig. 3A

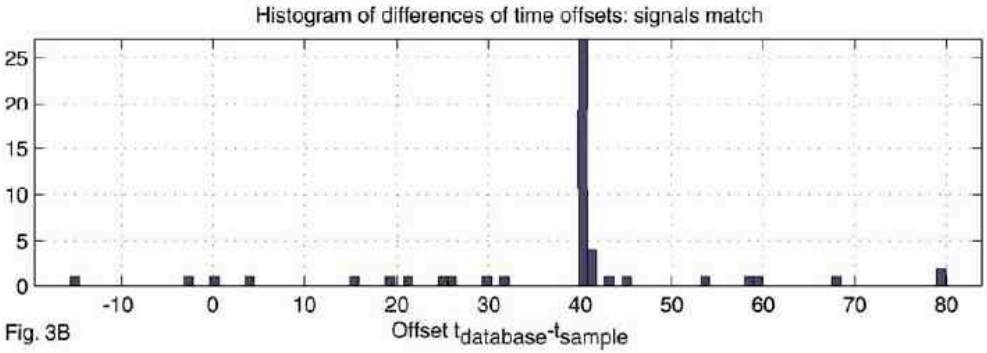


Fig. 3B

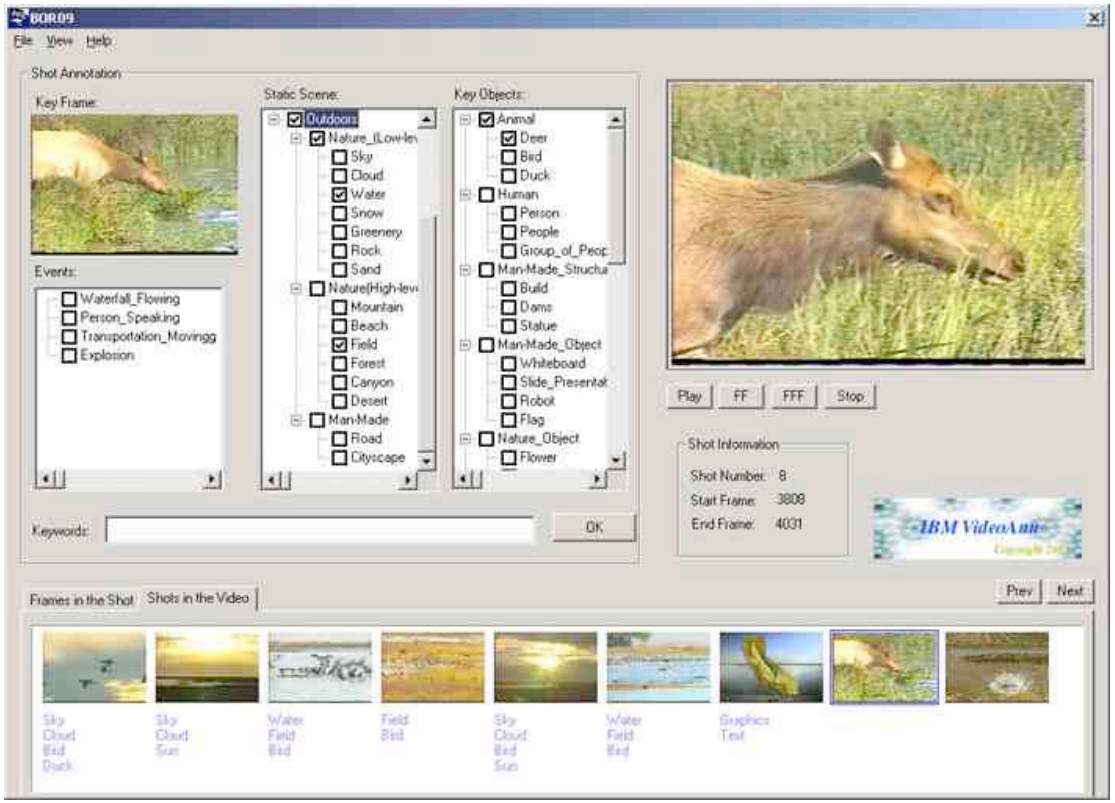
Example: Shazam Music Tagging (4)

- 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)

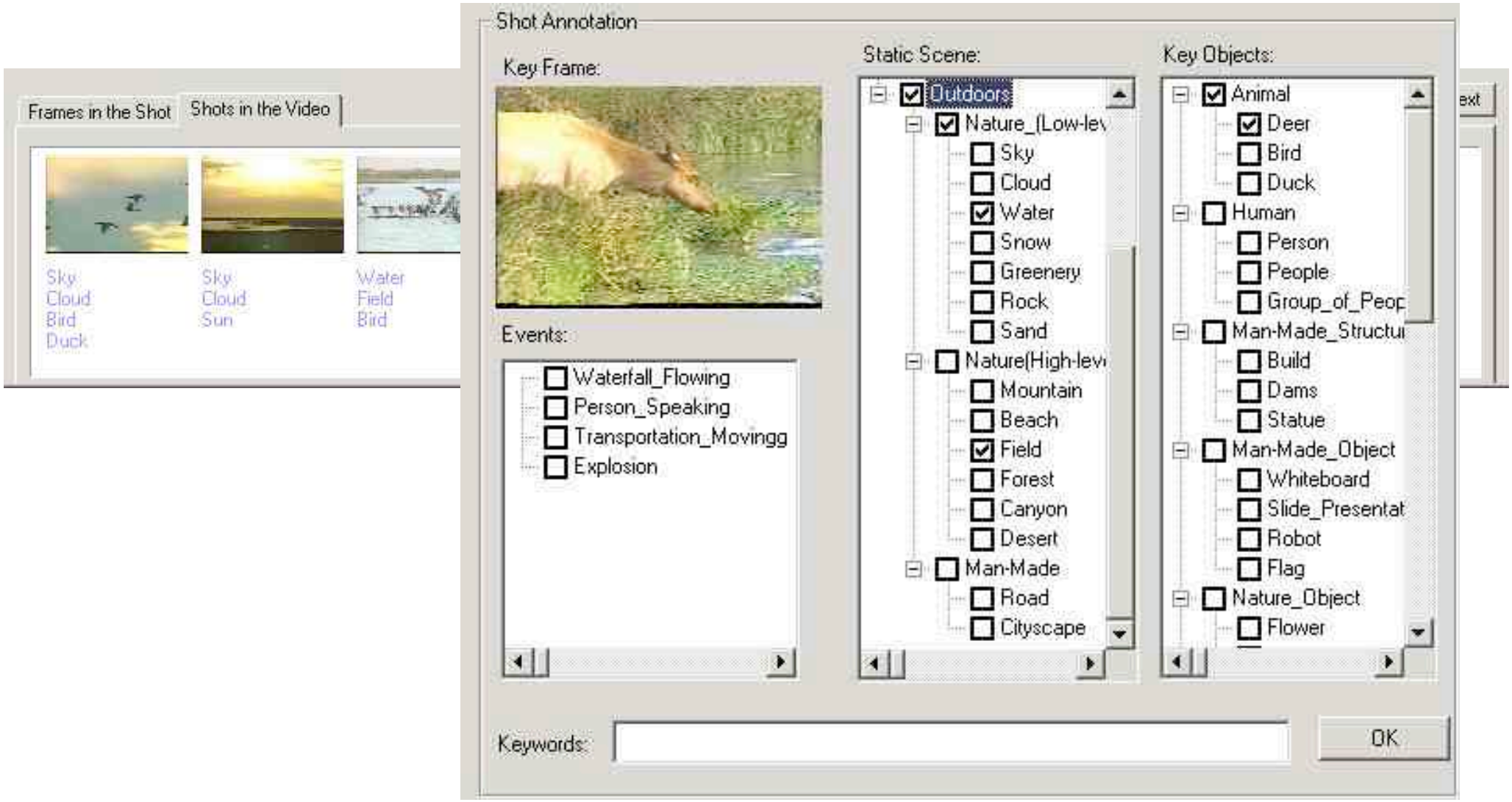
8 Multimedia Content Description

IBM VideoAnnEx (1)

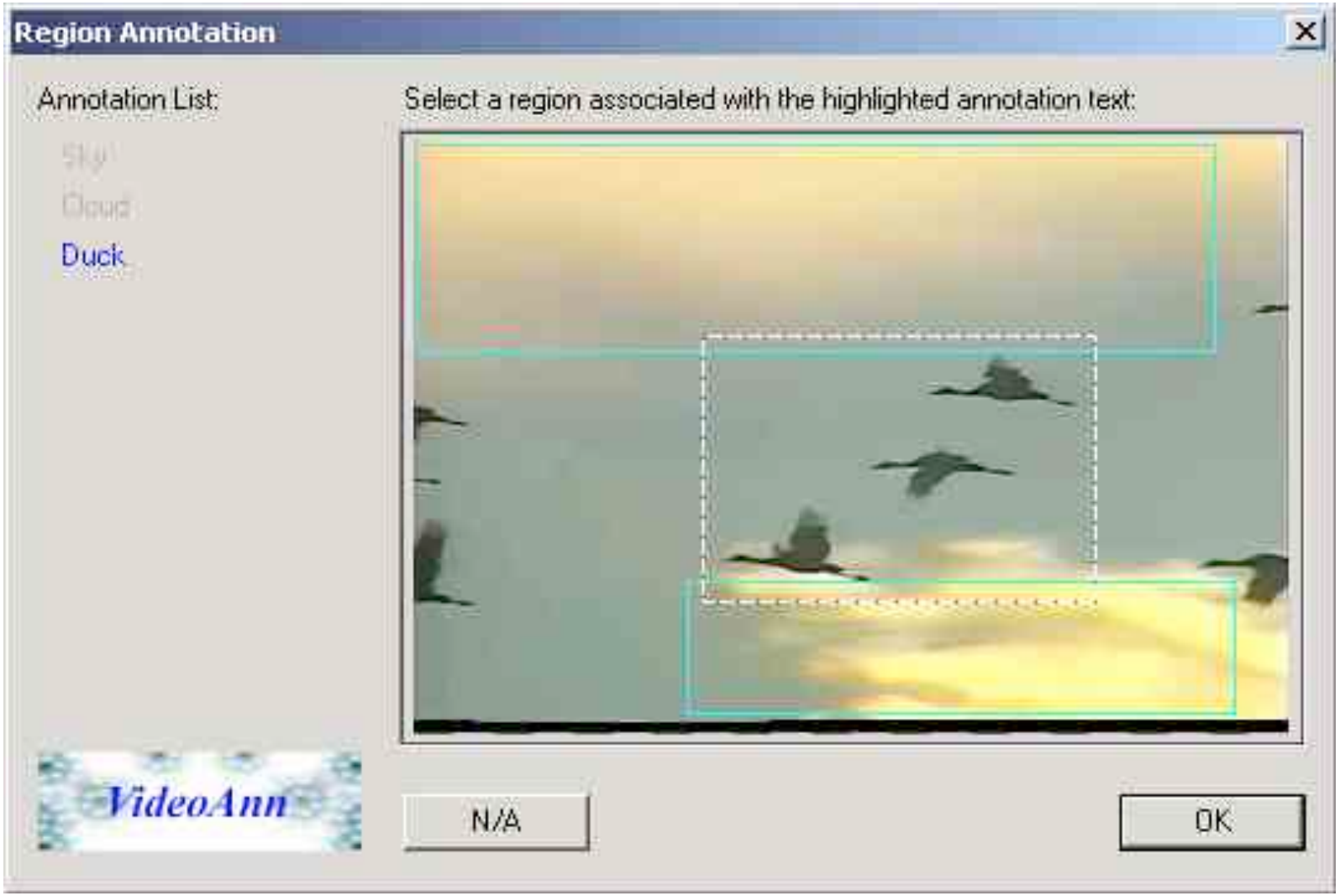
- Support tool for manual annotation of video sequences with MPEG-7 metadata
  - Experimental tool 2001-2003, no longer supported
  - Requires a basic lexicon of description items in addition to video file



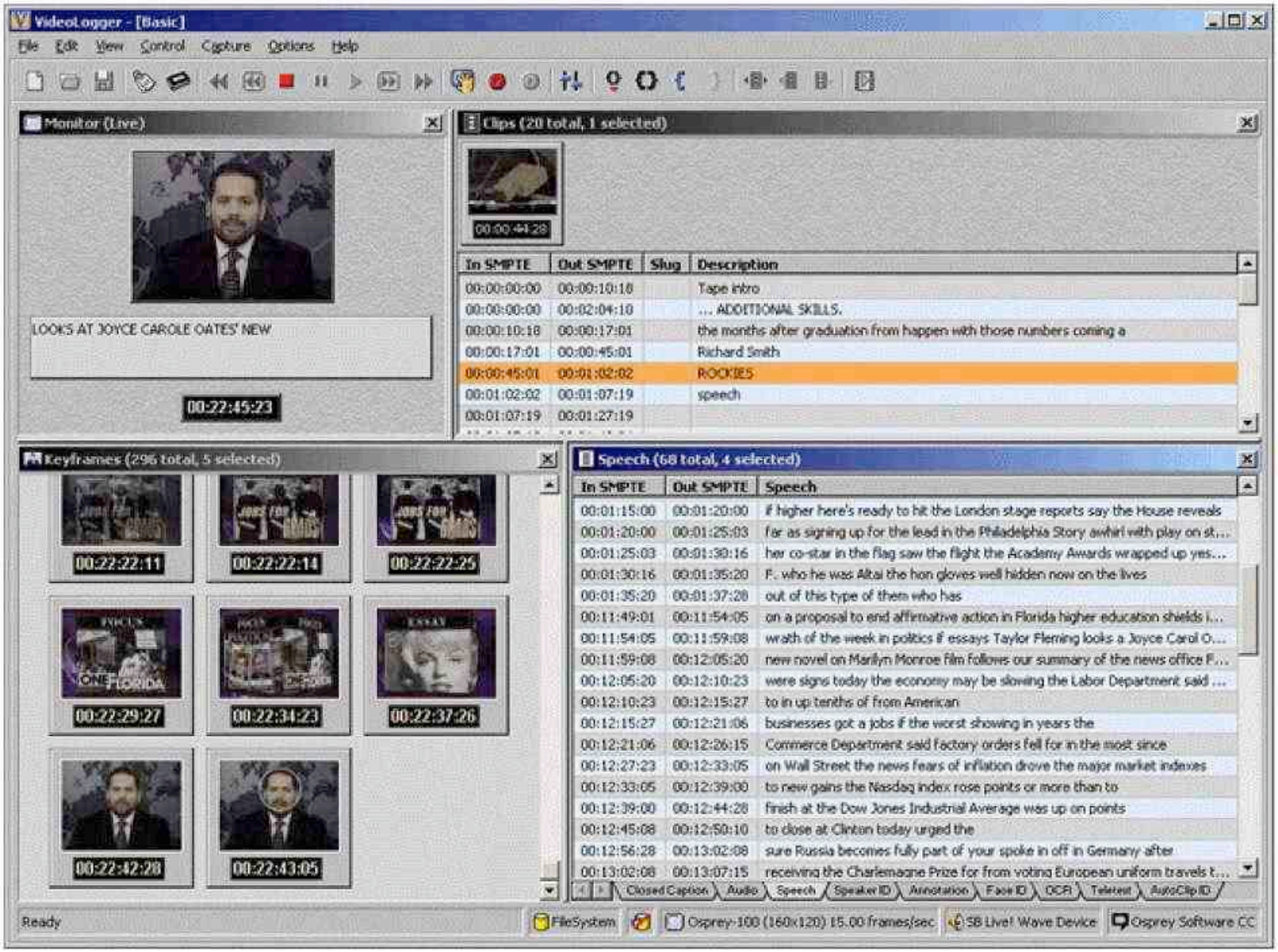
IBM VideoAnnEx (2)



IBM VideoAnnEx (3)



Virage VideoLogger



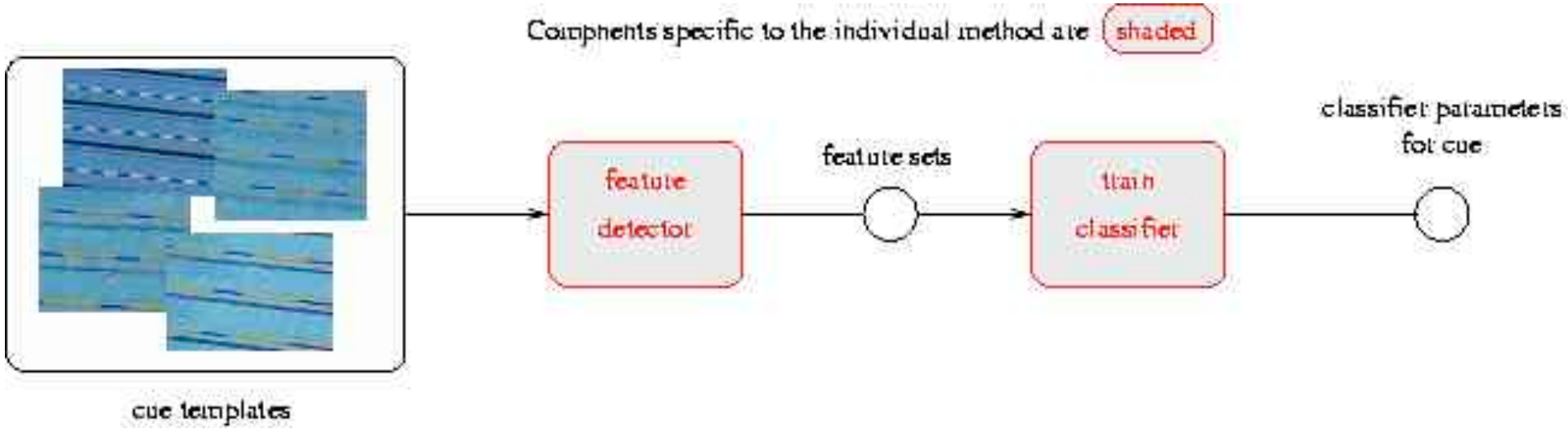
Techniques used by Virage VideoLogger

- Signal analysis algorithms to generate key frames for visual overview
- Speech-to-text transcription
- Sound identification
- Speaker identification
  - voice identification and face identification
- Analysis of embedded textual information:
  - close captioning, teletext
- External metadata:
  - PowerPoint presentations
  - EDLs
  - GPS data
  - transcripts
- Manual annotation:
  - Effective user interface (hot keys etc.)

Example: ASSAVID Sports Analysis System (1)

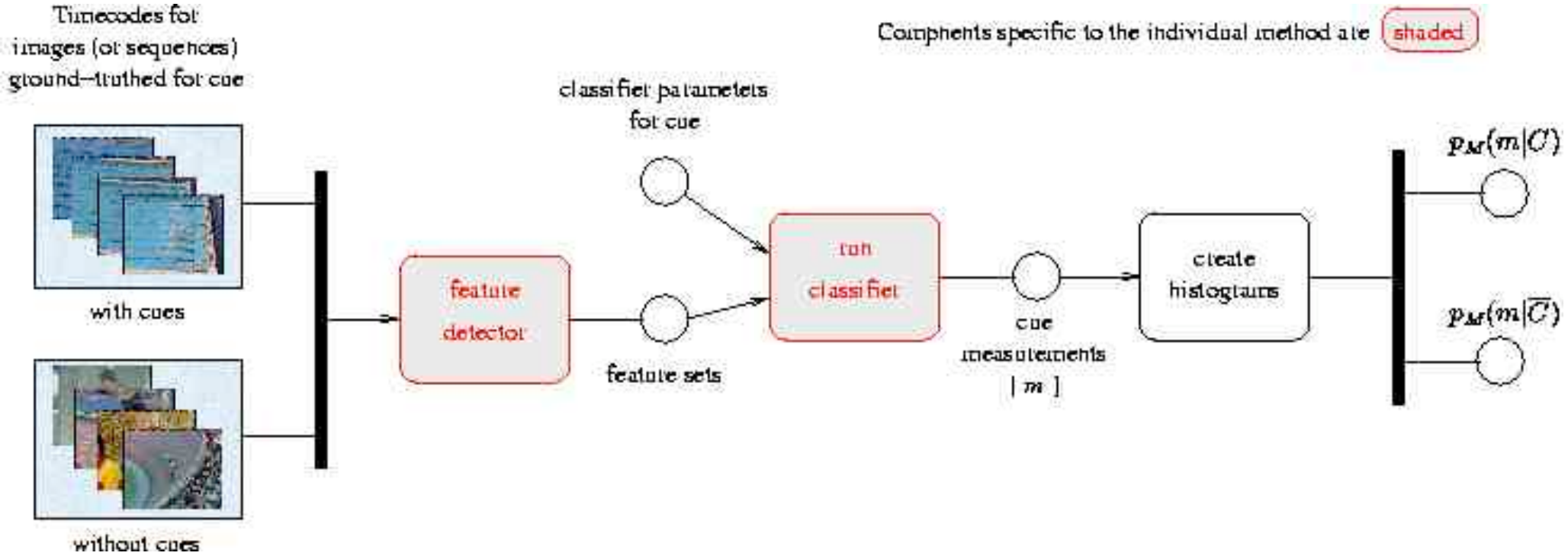
http://www.ee.surrey.ac.uk/CVSSP/SignalProcessing/Analysis/ASSAVID

Stage 1: Training



Example: ASSAVID Sports Analysis System (2)

Stage 2: Feature generation



Example: ASSAVID Sports Analysis System (3)

Stage 3: Testing