5 Multimedia Content Description

5.1 Metadata: Concepts and Overview

5.2 RDF: XML Metadata

5.3 Metadata for Authoring: AAF & SMPTE Standards

5.4 Generic Metadata Framework: MPEG-7

5.5 Advanced Multimedia Metadata in MPEG-7

5.6 Metadata for Music Information Retrieval

5.7 Automation of Video Metadata Extraction

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

- Feature:
  - Distinctive characteristic of the data which siginifies 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

Metadata Classification in MPEG-7

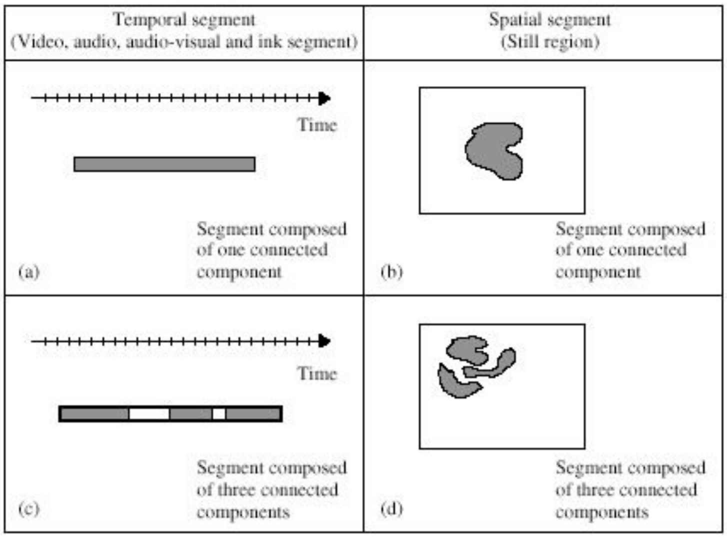
- Content Management
  - – Media information (e.g. file name, format, resolution)
  - – Creation information (e.g. creator, location, date)
  - – Usage information (e.g. rights)
- Content Structure
  - – Segments
  - – Segment relations
- Content Semantics

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

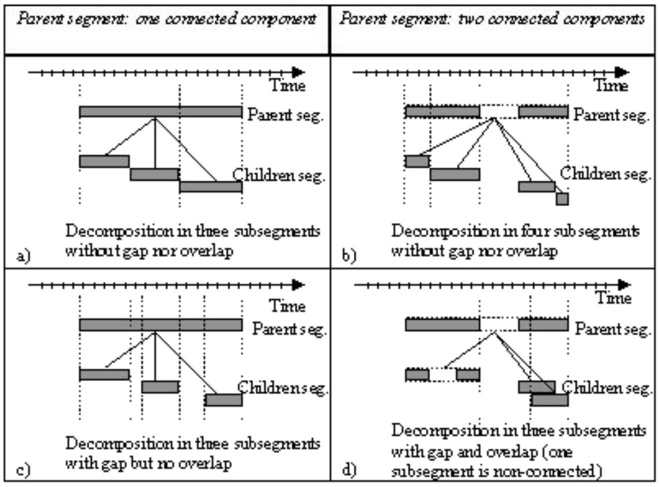

Temporal segment
(Video, audio, audio-visual and ink segment)
Spatial segment
(Still region)
Time
Segment composed of one connected component
(a)
Segment composed of one connected component
(b)
Time
Segment composed of three connected components
(c)
Segment composed of three connected components
(d)

Segment Attributes

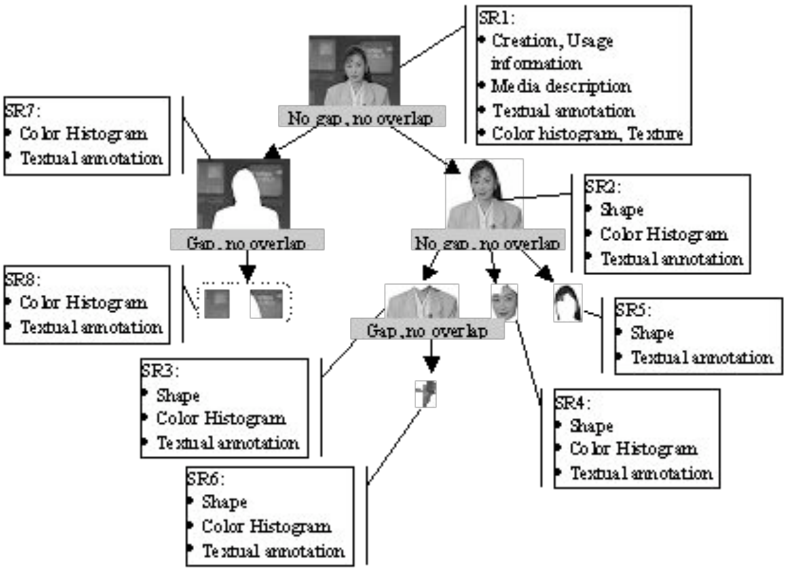
- Generic features
  - (media information, creation information, usage information, annotations)
- Media type dependent features:
  - (visual features, audio features)
- Specific features for segments
  - Mask Descriptor
    - » Spatial mask, Temporal mask, Spatio-temporal mask
  - Importance of descriptors
    - » 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


Parent segment: one connected component
Parent segment: two connected components
Time
Parent seg.
Children seg.
a)
Decomposition in three subsegments without gap nor overlap
b)
Decomposition in four subsegments without gap nor overlap
c)
Decomposition in three subsegments with gap but no overlap
d)
Decomposition in three subsegments with gap and overlap (one subsegment is non-connected)

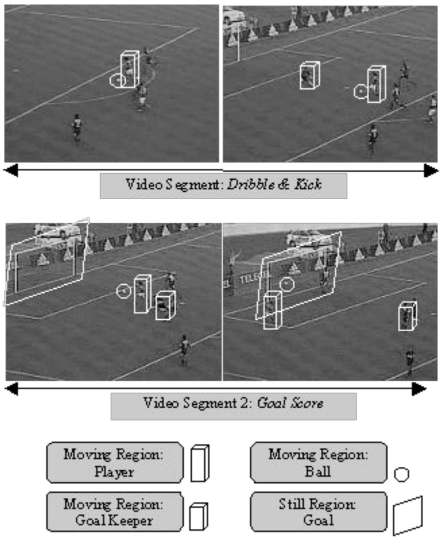
Example of Image Description

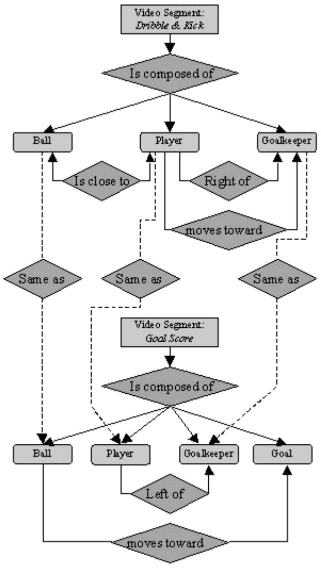


Structural Relations of Segments

- Content structure:
  - Either hierarchical segment decomposition
  - Or general segment relationship graph
- Predefined structural relations in MPEG-7 (can be extended):
  - 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
  - Generic:
    - Identical, union, disjoint
- For each relation, the inverse relation is implicitly defined.

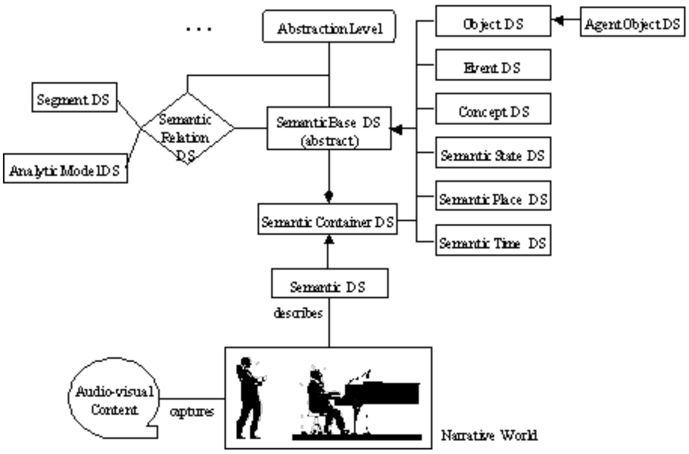
Video Segmentation with Moving Regions


Video Segment: *Dribble & Kick*
Video Segment 2: *Goal Score*
Moving Region: Player
Moving Region: Ball
Moving Region: Goal Keeper
Still Region: Goal

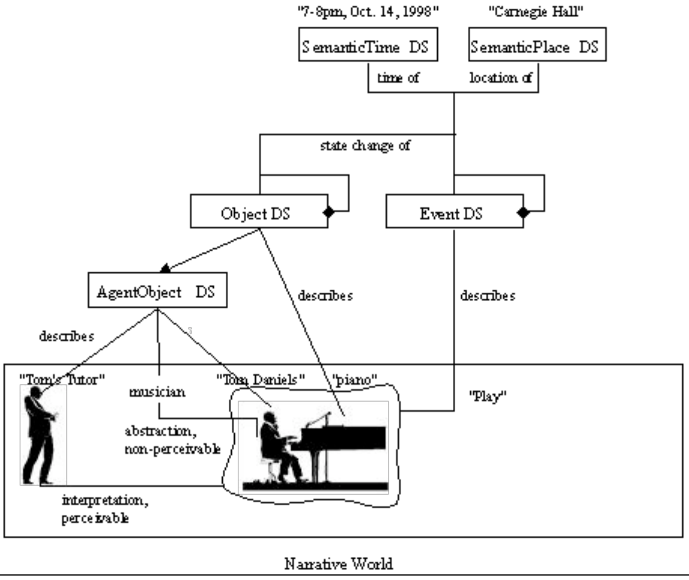


Content Semantics in MPEG-7

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

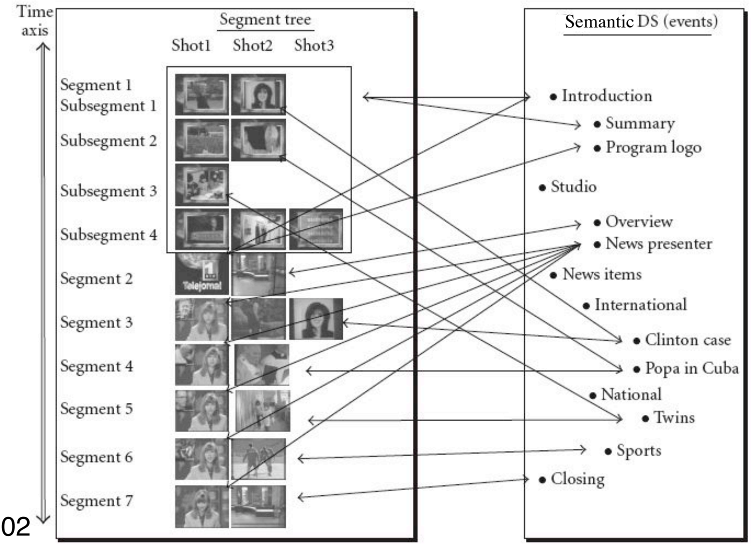


Content Semantics in MPEG-7: Example



Relating Structure and Semantics: Example

- Creating a semantic index for a video sequence

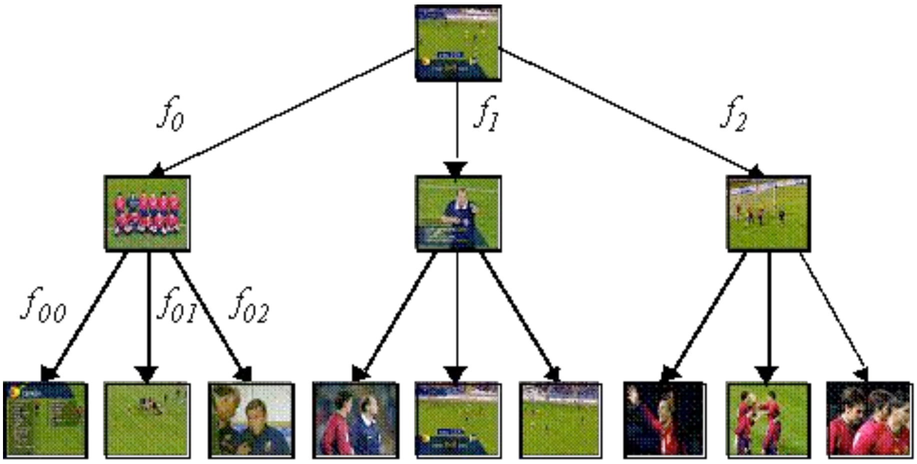


Salembier 2002

Navigation and Access

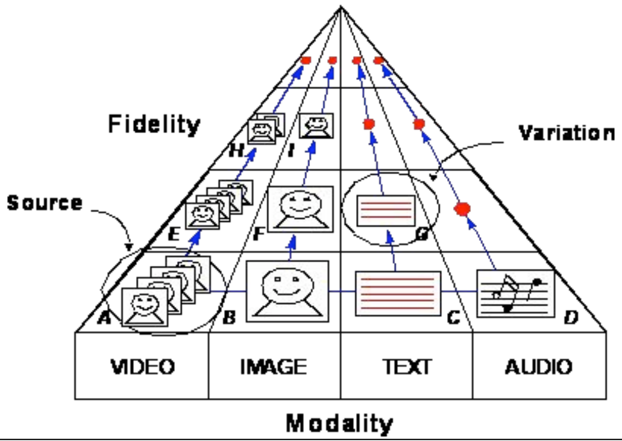
- Description schemes to facilitate navigation and access of audio-visual content:
  - Summaries
    - » Browsing, navigation, discovery, visualization, sonification
  - Views and partitions
    - » Representations in multiple domains, resolutions
  - Variations
    - » Different versions adapted to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

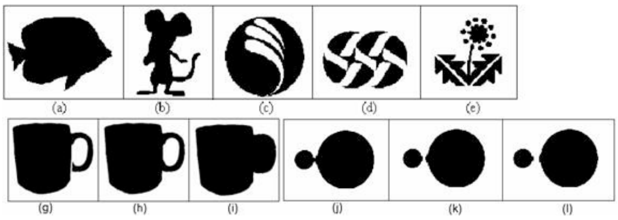
- Usage of compressed data formats:
  - E.g. frequency space transformation (JPEG), motion estimation (MPEG-2) can be re-used for feature extraction

Shape Descriptors

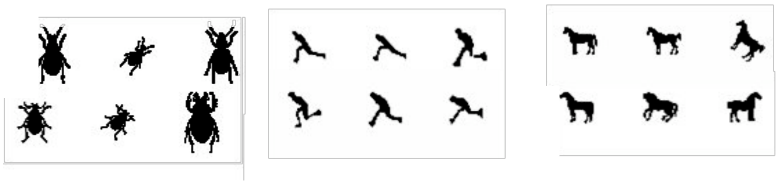
- Region shapes
  - Pixel distribution, using both boundary and internal pixels
  - Can describe complex objects with multiple disconnected regions
  - Shape analysis based on moments
    - » Angular Radial Transformation (ART)
- Contour shapes
  - Based on Curvature Scale-Space (CSS) representation 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:

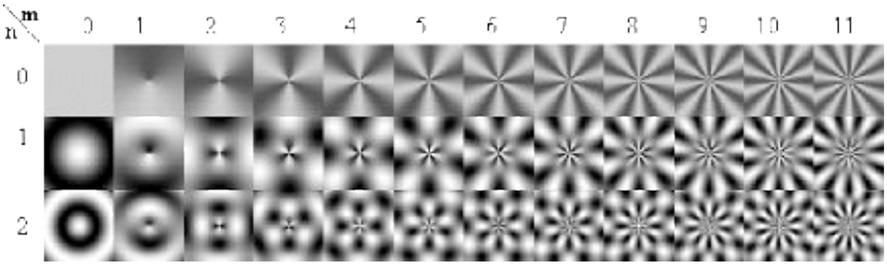


Contour shapes:



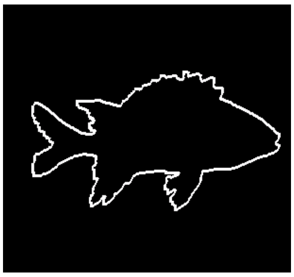
Angular Radial Transformation (ART)

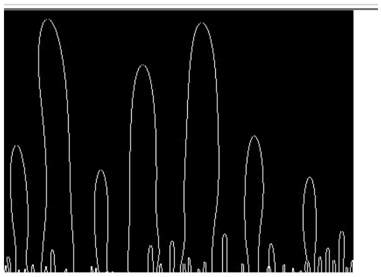
- Convert image information into angular and radial parts
- Represent image as coefficients of basis functions
- First 36 basis functions:



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-cronssings of various scales
- http://www.ee.surrey.ac.uk/Research/VSSP/imagedb/affine.html

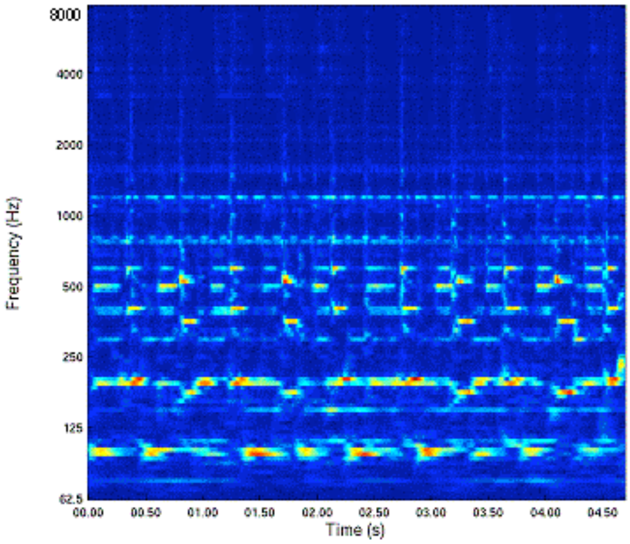




MPEG-7 Audio Description Tools

- 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-levell audio descriptors:
  - Sound recognition and indexing
  - Musical instrument timbre description
  - Melody description tools
  - Spoken language recognition

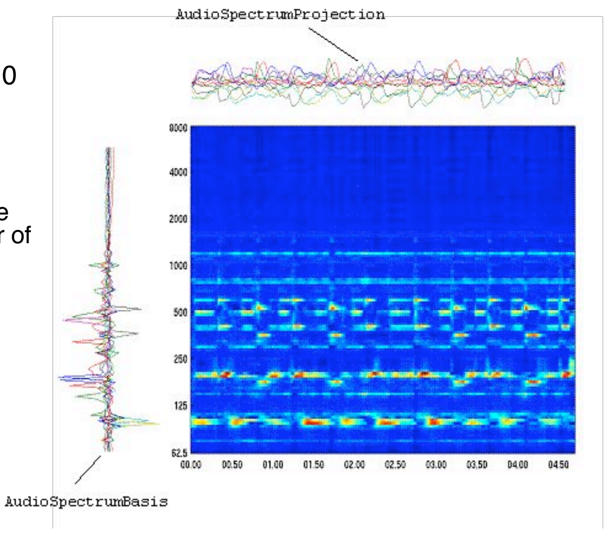
Spectral Analysis with AudioSpectrumEnvelope



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



5 Multimedia Content Description

5.1 Metadata: Concepts and Overview

5.2 RDF: XML Metadata

5.3 Metadata for Authoring: AAF & SMPTE Standards

5.4 Generic Metadata Framework: MPEG-7

5.5 Advanced Multimedia Metadata in MPEG-7

5.6 Metadata for Music Information Retrieval

5.7 Automation of Video Metadata Extraction

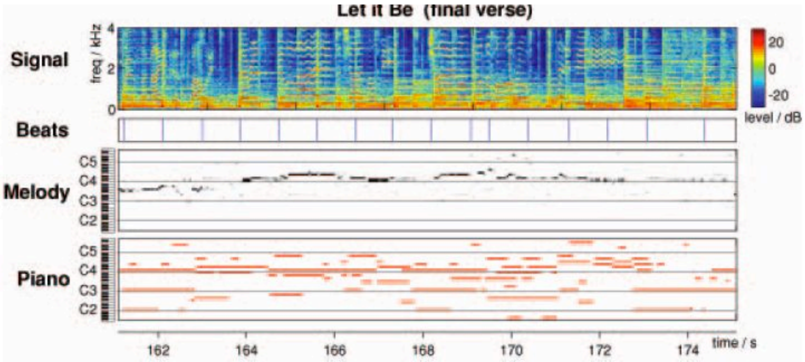
Literature:

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”)

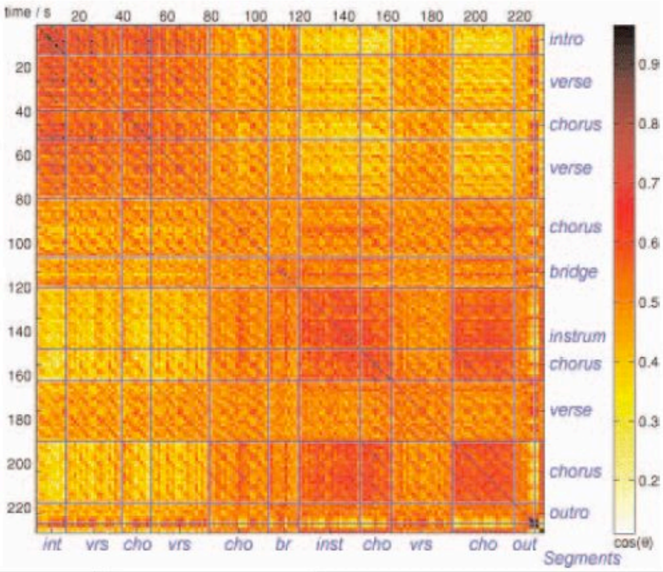
Automatic Score Transcription



- 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

- Commercial service for mobile phones: Identify music from a short audio sample *(query by example)*
  - See http://www.shazam.com
- 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)
  - Few “anchor” peaks are combined with peaks in a certain surrounding zone (time and frequency offsets)
    - » Combinatorial hashing creates 32b fingerprint hash token
  - Temporal alignment greatly accelerates matching process
- Real system:
  - “a few dozen” x86-based servers with 8GB of RAM, Linux
  - June 2006: More than three million tracks

5 Multimedia Content Description

5.1 Metadata: Concepts and Overview

5.2 RDF: XML Metadata

5.3 Metadata for Authoring: AAF & SMPTE Standards

5.4 Generic Metadata Framework: MPEG-7

5.5 Advanced Multimedia Metadata in MPEG-7

5.6 Metadata for Music Information Retrieval

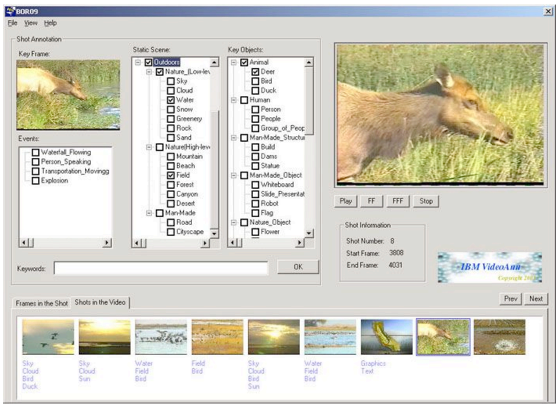
5.7 Automation of Video Metadata Extraction

Literature:

www.virage.com

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



Events:
- ☐ Waterfall_Flowing
- ☐ Person_Speaking
- ☐ Transportation_Movingg
- ☐ Explosion

Static Scene:
- ☑ Outdoors
  - ☑ Nature_(Low-lev
    - ☐ Sky
    - ☐ Cloud
    - ☑ Water
    - ☐ Snow
    - ☐ Greenery
    - ☐ Rock
    - ☐ Sand
  - ☐ Nature(High-lev
    - ☐ Mountain
    - ☐ Beach
    - ☑ Field
    - ☐ Forest
    - ☐ Canyon
    - ☐ Desert
  - ☐ Man-Made
    - ☐ Road
    - ☐ Cityscape

Key Objects:
- ☑ Animal
  - ☑ Deer
  - ☐ Bird
  - ☐ Duck
- ☐ Human
  - ☐ Person
  - ☐ People
  - ☐ Group_of_Peop
- ☐ Man-Made_Structu
  - ☐ Build
  - ☐ Dams
  - ☐ Statue
- ☐ Man-Made_Object
  - ☐ Whiteboard
  - ☐ Slide_Presentat
  - ☐ Robot
  - ☐ Flag
- ☐ Nature_Object
  - ☐ Flower

Keywords:

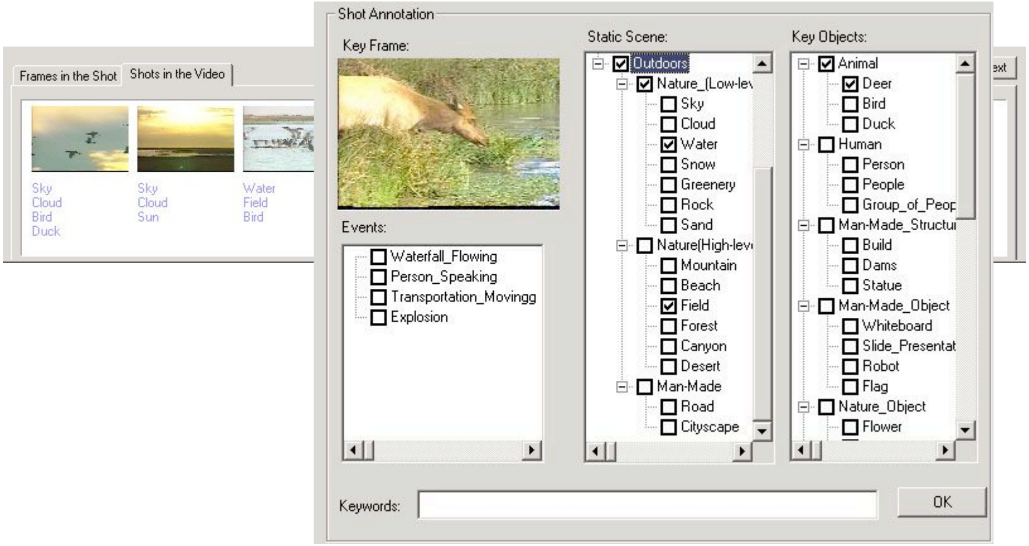
Shot Information

Shot Number: 8

Start Frame: 3808

End Frame: 4031

IBM VideoAnnEx (2)



Frames in the Shot | Shots in the Video

Sky
Cloud
Bird
Duck

Sky
Cloud
Sun

Water
Field
Bird

Shot Annotation

Key Frame:

Events:

- □ Waterfall_Flowing
- □ Person_Speaking
- □ Transportation_Movingg
- □ Explosion

Static Scene:

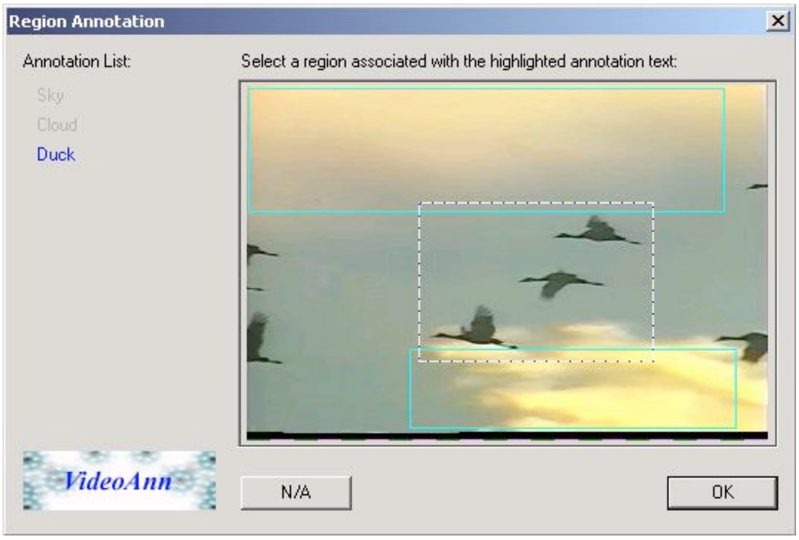
- ☑ Outdoors
  - ☑ Nature_(Low-lev
    - □ Sky
    - □ Cloud
    - ☑ Water
    - □ Snow
    - □ Greenery
    - □ Rock
    - □ Sand
  - □ Nature(High-lev
    - □ Mountain
    - □ Beach
    - ☑ Field
    - □ Forest
    - □ Canyon
    - □ Desert
  - □ Man-Made
    - □ Road
    - □ Cityscape

Key Objects:

- ☑ Animal
  - ☑ Deer
  - □ Bird
  - □ Duck
- □ Human
  - □ Person
  - □ People
  - □ Group_of_Peop
- □ Man-Made_Structur
  - □ Build
  - □ Dams
  - □ Statue
- □ Man-Made_Object
  - □ Whiteboard
  - □ Slide_Presentat
  - □ Robot
  - □ Flag
- □ Nature_Object
  - □ Flower

Keywords:

IBM VideoAnnEx (3)



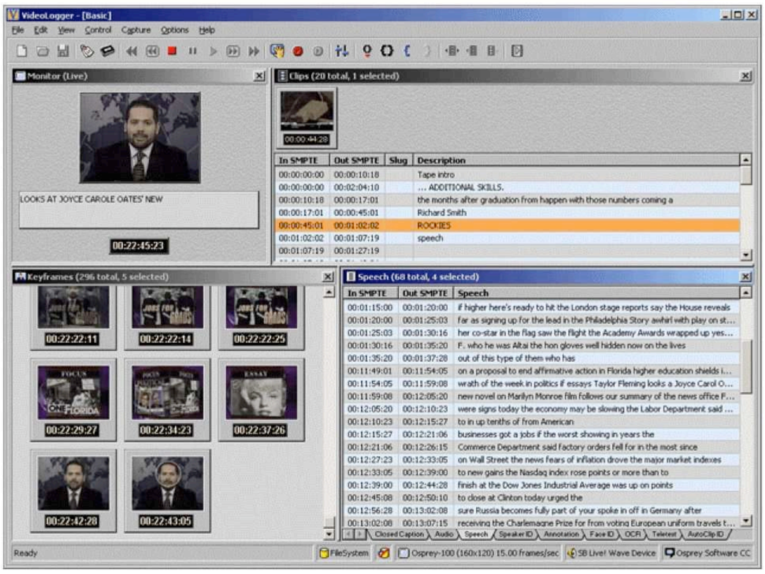
Region Annotation

Annotation List:

- Sky
- Cloud
- Duck

Select a region associated with the highlighted annotation text:

Virage VideoLogger



LOOKS AT JOYCE CAROLE OATES' NEW

In SMPTE	Out SMPTE	Slug	Description
00:00:00:00	00:00:10:18		Tape intro
00:00:00:00	00:02:04:10		… ADDITIONAL SKILLS.
00:00:10:18	00:00:17:01		the months after graduation from happen with those numbers coming a
00:00:17:01	00:00:45:01		Richard Smith
00:00:45:01	00:01:02:02		ROCKIES
00:01:02:02	00:01:07:19		speech
00:01:07:19	00:01:27:19		

In SMPTE	Out SMPTE	Speech
00:01:15:00	00:01:20:00	if higher here's ready to hit the London stage reports say the House reveals
00:01:20:00	00:01:25:03	far as signing up for the lead in the Philadelphia Story awhirl with play on st…
00:01:25:03	00:01:30:16	her co-star in the flag saw the flight the Academy Awards wrapped up yes…
00:01:30:16	00:01:35:20	F. who he was Altai the hon gloves well hidden now on the lives
00:01:35:20	00:01:37:28	out of this type of them who has
00:11:49:01	00:11:54:05	on a proposal to end affirmative action in Florida higher education shields i…
00:11:54:05	00:11:59:08	wrath of the week in politics if essays Taylor Fleming looks a Joyce Carol O…
00:11:59:08	00:12:05:20	new novel on Marilyn Monroe film follows our summary of the news office F…
00:12:05:20	00:12:10:23	were signs today the economy may be slowing the Labor Department said …
00:12:10:23	00:12:15:27	to in up tenths of from American
00:12:15:27	00:12:21:06	businesses got a jobs if the worst showing in years the
00:12:21:06	00:12:26:15	Commerce Department said factory orders fell for in the most since
00:12:27:23	00:12:33:05	on Wall Street the news fears of inflation drove the major market indexes
00:12:33:05	00:12:39:00	to new gains the Nasdaq index rose points or more than to
00:12:39:00	00:12:44:28	finish at the Dow Jones Industrial Average was up on points
00:12:45:08	00:12:50:10	to close at Clinton today urged the
00:12:56:28	00:13:02:08	sure Russia becomes fully part of your spoke in off in Germany after
00:13:02:08	00:13:07:15	receiving the Charlemagne Prize for from voting European uniform travels t…

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.)