ESSIR 2003

Models, tools for Video

Georges Quénot

CLIPS-IMAG CNRS

Models and tools for Video

- Media Specificity
- Segmentation
- Indexing and Retrieval
- MPEG-7
- TREC
- Research Systems
- Commercial Systems
- Web Systems
- Conclusion

Media specificity

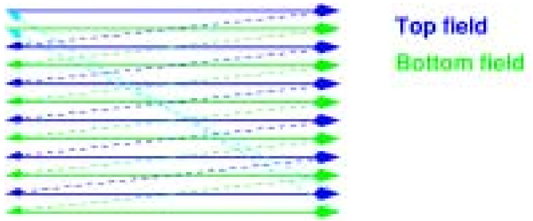
- Coherent and synchronized combination of simpler media
  - Image (animated)
  - Audio
  - Text (closed captions, subtitles)
  - Others (Virtual Reality, ...)
  - *Lots of* formats, resolutions and qualities
- Stream aspect
  - Temporal structure
- *Passage* retrieval from several media
- Various levels of semantic gap and uncertainties
- Contextual information

Animated image

- Interlaced and progressive schemes: images and fields
  - Image/field size
  - Image/field rate
- Compression
  - Image by image: spatial redundancy
  - Image sequence: spatial and temporal redundancies
  - From 28 kbit/s (176 x 144 x 12.5 Hz, sequence coding, realmedia) up to 28 Mbit/s (720 x 288 x 50 Hz, interlaced image coding, DV) or 270 Mbit/s (704 x 288 x 50 Hz, uncompressed, 10-bit D1).
- Very variable quality

Image interlacing

- Compromises resolution versus continuity
- *Complicates* image sequence processing
- Apply only to “full resolution” streams



Audio

- From 8 kHz up to 48 kHz sample rates
- Mono, stereo or more (5.1, ...)
- Multiple audio streams (multilingual DVDs)
- Compressed from ~5 kbit/s up to 256 kbit/s (1.41 Mbit/s for uncompressed, audio CD) for stereo streams
- Very variable quality

Text

- Subtitles or source speech transcripts
  - Good quality text inserted as separate streams
  - Accurate and relevant
  - When available...
- + Closed captions
  - Text actually contained in the image stream
  - Hard to recover, high miss and error rates
- + Field text (text in the scene)
  - Even worse...
- + Automatic Speech Recognition
  - Text actually contained in the audio stream
  - Significant word error rate but usable

Other stream types

- Animated Artefacts streams (MPEG-4)
  - Audio, text and image substreams
  - May contain accurate and relevant semantic information
  - Could be used of indexing and retrieval
  - Currently exotic and not widely used
- Musical Instrument Digital Interface streams
- Speech synthesis data streams
- ...

Applications

- Search
  - General (web, ...)
  - Domain specific (medical, military, ...)
- Filtering
  - Technological (or political, ...) survey
  - Offending content
- Metrics
  - Commercials impact estimation (spots, logos)
- Copyright check
- Domains
  - Archives, web
  - Conference, meetings

Video Segmentation

- Image Segmentation
  - Shot segmentation,
  - Object segmentation and tracking,
  - Camera and object motion,
  - Key frame extraction
- Audio Segmentation
  - Silence / music / noise / speech / ...
  - Male / female
  - High quality / telephone / ...
  - Speaker / known speaker
- Sub-shot (micro-) segmentation
- Story / topic (macro-) segmentation

Shot segmentation

- Direct image comparison or matching
  - Sum of square differences, ...
  - With or without motion compensation
- Descriptor extraction and comparison
  - Color momentums or histograms
  - Texture, shapes
- Other methods
  - Rough contour tracking
- Compressed domain methods
  - DC image comparison
  - Motion vectors, ratio of forward versus backward vectors
  - Bit rate variation, ratio of predicted versus intra-coded blocks
  - Specific to media encoding, moderate quality

Shot segmentation

- "Cuts" versus gradual transitions
- Separated methods:
  - Cuts: search for discontinuities in images or descriptors
  - Other: ad'hoc methods, specific searches for wipes, dissolves, block changes, ...
  - Plus: photographic flashes filtering, ...
  - Need for detector outputs' fusion
- Integrated multi-resolution methods:
  - Filtering of descriptors time derivatives at various time scales
  - Search for peaks with sophisticated filtering:
    - » Peak location -> transition (center) location
    - » Scale of the highest peak -> transition duration

Object segmentation and camera motion

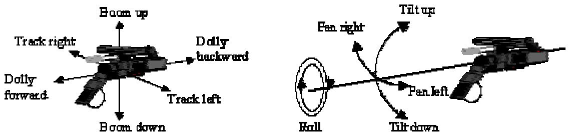
- Classical still image object segmentation (using color, texture, regions, contours, ...) plus:
- Extraction of objects moving relatively to a background and of relative camera/background motion:
  - Objects are identified as not following the background motion
  - Background motion is identified by applying a motion model to the part of the image excluding mobile objects
  - Reciprocal dependency can be broken using iterative methods if the background occupy a large enough part of the images
  - Several alternative motion models have to be considered
  - Speed versus accuracy compromises (MPEG vectors versus full optical flow computation)
- Combination of still image and motion based object segmentation
- Bonus: mosaic images or 3D views of the background

Iterative simultaneous extraction of mobile objects and of relative camera motion

- Start with assumption of no mobile objects
- Estimate motion parameter from whole image
- Estimate probability of belonging to the background for each pixel
- Re-estimate motion parameter from pixels with a high probability of belonging to the background
- Re-estimate probability of belonging to the background for each pixel
- Iterate 4. - 5. until convergence or given count

- No need to make any binary decision
- Provides camera motion and background segmentation

Background / camera motion models 7 degrees of freedom (MPEG-7 descriptors)



Parametric camera model, 7 parameters:

- Translation: track, boom, dolly,
- Rotation: tilt, pan, roll,
- Zoom or focal length

One- or two-level camera motion search

- Direct camera motion search or:
- Camera motion search from a previously extracted image motion model
- Parametric image motion models:

Geometric Model	Params	x'	Y'
Translational	**2**	x+a	y+b
Similitude	**4**	ax-by+c	bx+ay+d
Affine	**6**	ax+by+c	dx+ey+f
Homographic	**8**	(ax+by+c)/(gx+hy+1)	(dx+ey+f)/(gx+hy+1)
Quadratic	**12**	ax^2+by^2+cxy+dx+ey+f	gx^2+hy^2+ixy+jx+ky+l

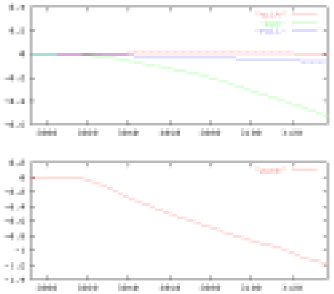
Techniques for Motion Field Extraction

Method	Type	Description
Dense Optical Flow Calculation	**Pixel or Sub Pixel Level** **(very slow)**	*Exhaustive pixel correspondance search between frames giving a dense motion field for each pixel*
Phase Correlation Motion Estimation	**Block Based** **(relatively fast)**	*Using the shift property of the Fourier Transform, fairly precise motion of image blocks can be calculated with acceptable speed*
Block Matching for Intensity Difference	**Block Based** **(fast, depending on search strategy)**	*Motion fields extracted on a block level by finding the best maching area that minimizes the total intensity difference (squared or absolute)*
Readymade Motion Vectors **(eg. MPEG)**	**Block Based** **(available)**	*Assuming the motion vectors represent the approximate intensity flow, they can be read directly from the MPEG stream and used*

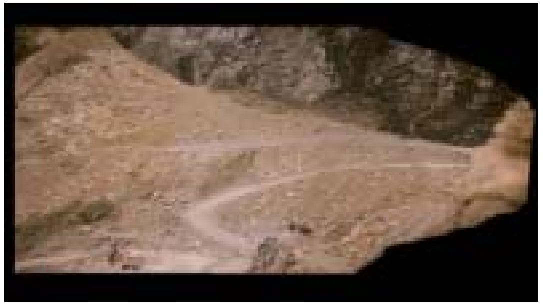
Background / camera relative motion search

- No background motion (simple check)
- Motion without parallax (two-level search):
  - Rotations (3) and focal length,
  - Search for an homographic transform,
  - Mosaicing (panoramic view) and mobile objects,
- Motion with parallax (one-level search):
  - Rotations (3), translations (3) and focal length,
  - « Motion and structure from motion », « paraperspective decomposition » method from Poelman et Kanade (1993).
  - Three-dimensional view of the backgroud
- Irregular motion (crowd, waves, ...)

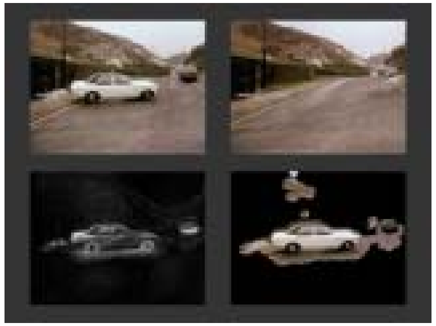
Camera motion, aim1mb08.mpg document, 2992-3137 sequence



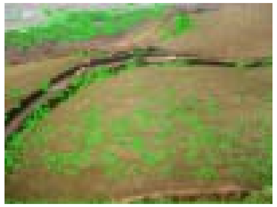
Panoramic view, aim1mb08.mpg document, 2992-3137 sequence



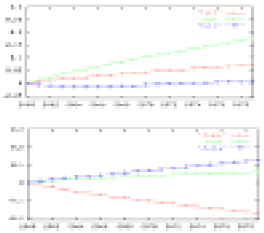
Mobile objects segmentation and tracking aim1mb08.mpg, 1671-1715 sequence



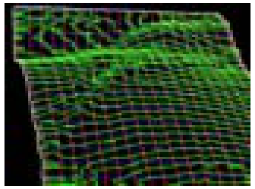
Motion with parallax: Tracking of feature points aim1mb08.mpg, 2860-2879 sequence



Motion with parallax



Three-dimensional scene structure



Key frame extraction

- Choice of one or more representative key frame per continuous shot
- First or center frame
- Frame with low motion, following a zoom, ...
- Frame with high contrast
- Weighted combination of criteria
- Panoramic view when available
- Frame showing a best view of an object (or frontal face)
- Multiple key frame selection according to content change detection and/or important feature detection
- Maximization of inter-shot key frame dissimilarity
- Used for content display (browsing) or indexing

Audio Segmentation

- Silence / music / noise / speech / male / female / high quality / telephone / speaker / known speaker / emotion / ...
- Feature extraction: spectral analysis (MFCC, LPC, plus ...) on 10-20 ms windows
- Class modeling or clustering using gaussian mixture densities
- Known classes (music, speech, male, anger, ...) or classes to be built (one for each unknown speaker)

Micro-segmentation

- Segmenting into units shorter than a shot
- May be hierarchical
- Change in background / camera relative motion
  - Start / stop of a zoom or a pan
- Change in object motion
  - Start, stop or direction change of objects
  - Appearance or desappearance of objects/persons
- Partial / local transitions
  - Small image appearance, desappearance or change
- Speaker or topic change
- Useful for key frames selection and content analysis

Macro-segmentation

- Segmenting into units longer than a shot
- May be hierarchical
- Not necessarily aligned with image or audio transitions
- Generally according to semantic changes like switch of topic within a TV journal
- Use of various clues:
  - Visual or audio jingles, black or blue frames,
  - Topic detection and tracking from audio transcription,
  - Pattern detection from audio transcript,
  - Detection of text or small image appearance or change.
- Useful for determining appropriate boundaries of retrieved passages

Content indexing and retrieval

- Low, intermediate and high level indexing
- Medium specific indexing
  - audio segment, image segment, text element
- Segment specific indexing
  - macro-segment indexing, fusion of medium specific indexing
- Topic / object / person / event specific indexing
- Multiple views with cross-references
- Possible use of conceptual graphs, ontologies and lists of individuals

Low level image segment indexing

- Color, texture, contour, shape, regions, points of interest, ... : exactly as for still image -> various classes of descriptors and of tools for extracting and comparing them, generally applied on selected key frames
- Plus: motion descriptors:
  - Global camera/background motion, quantitative or qualitative
  - Motion statistics : mean, standard deviation, entropy
  - Mobile objects count, sizes and motions
- Associated comparison / matching methods
  - Euclidian distance, cosinus measure, ...

High level image segment indexing

- Classification from low level image descriptors (indoors, cityscape, ...)
- Specific detection of human faces and bodies
  - Detection, classification, identification, facial expression
  - Still image techniques plus use of redundancy and motion
- Specific detection of definite objects (generally domain specific targets)
- Automatic processing currently limited:
  - Small number of classes (up to a few tens of classes)
  - Low recognition rates, degrading with number of classes
  - Notable exception: face detection and recognition
- Manual input required for professional quality

Audio segment indexing

- Semantic segmentation and classification
- Speaker recognition
- Automatic speech recognition
- NOT directly semantic indexing because of synonymy, polysemy and errors but quite close and usable especially when word sets are used for queries (redundancy and implicit desambiguation)
- Non linguistic features like emotion or prosody
- Music genre and noise classification

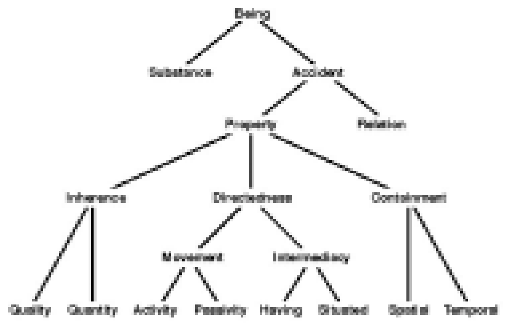
Macro segment indexing

- Fusion of medium specific descriptors using description schemes
- Links between related elements in different media descriptions (between who is seen and who speaks for instance)
- Organization of indexed elements : list, relation with ontologies and individual bases, links between them
- Automatic, semi-automatic or manual process
- Can be done at various levels of hierarchy and consistently across the hierarchy

Topic / object / person / event indexing

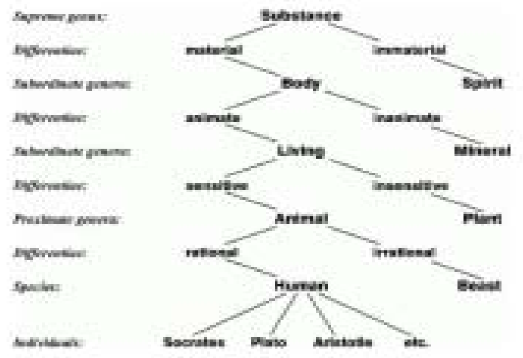
- Inverse table from indexed segments
- Part of ontology and individual bases
  - Ontology : network of concepts with typed links (hyponyms, hyperonyms, meronyms, ...) can be general or domain specific
  - Individual bases : lists of persons, cities, countries, institutions, possibly linked to ontology concepts
  - Knowledge bases: help to create additional links
- Linked to macro segments, medium segments and/or micro-segments
- Links can be typed, for instance when a person is linked to a medium segment, this person can be seen, heard, talked of, ...
- Relations between elements can be represented using conceptual graphs

Aristotle’s Ontology

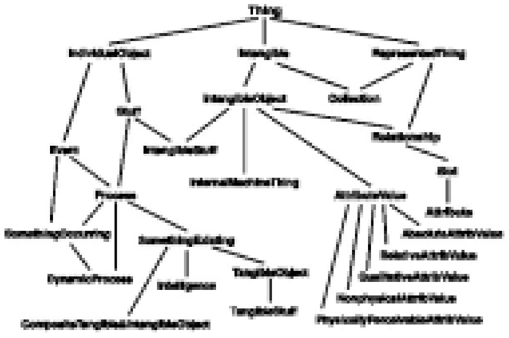


Aristotle's categories as arranged by Franz Brentano

Tree of Porphyry (as drawn by Peter of Spain)

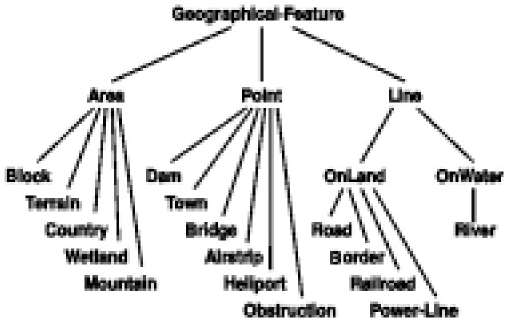


Largest Formal Ontology (John F. Sowa)



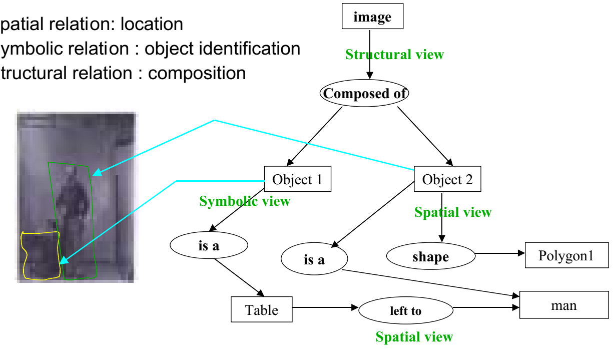
Top-level categories used in Cyc

Small Hand-Coded Ontologies



Extended Model for Image Retrieval

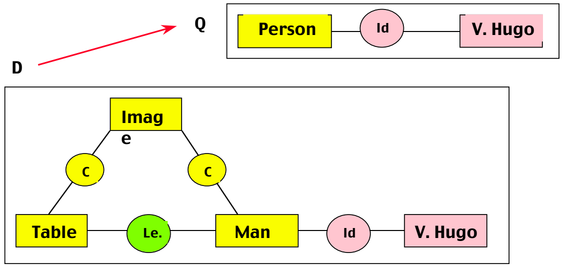
- Spatial relation: location
- Symbolic relation : object identification
- Structural relation : composition



EMIR_ [Mechkour [95] (laboratoire Clips-IMAG)]

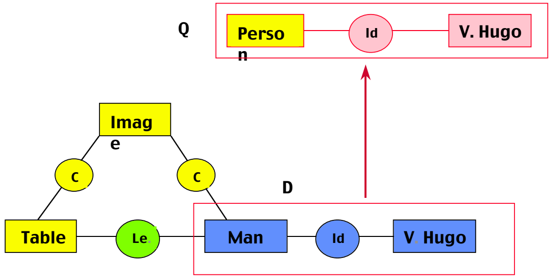
Matching: an example

- Theoretical model: “D implies Q”
  - First order logic

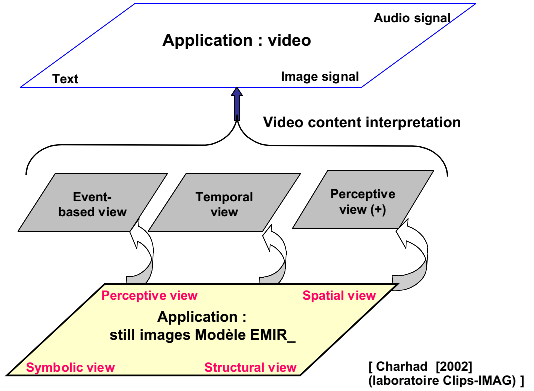


Projection: an exemple

- Operational model:
  - Graph projection

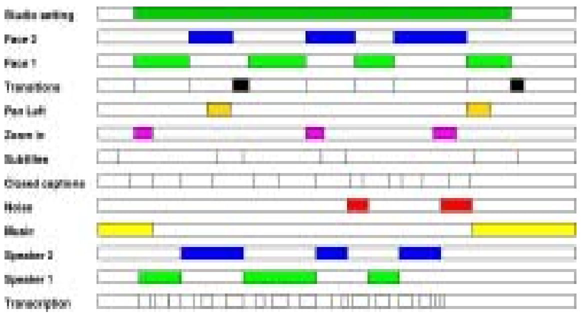


Extended Model for Video Retrieval

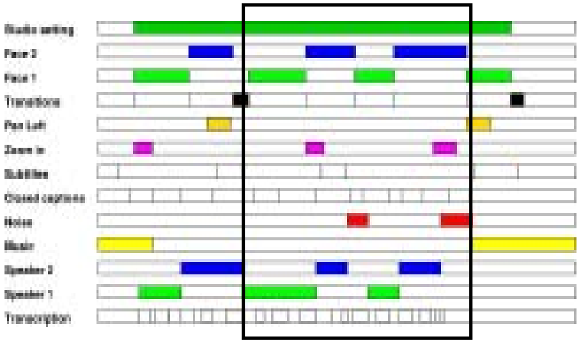


[Charhad [2002]
(laboratoire Clips-IMAG)]

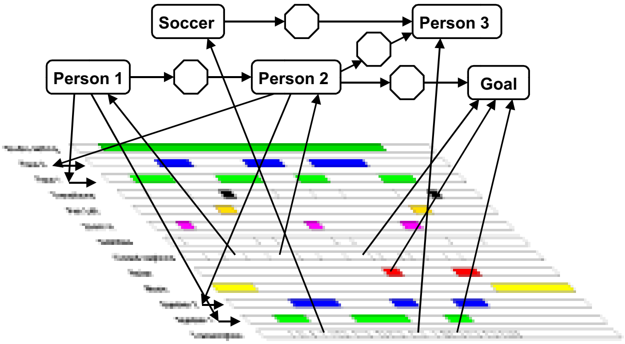
Track indexing



Relevant segment



Semantic indexing

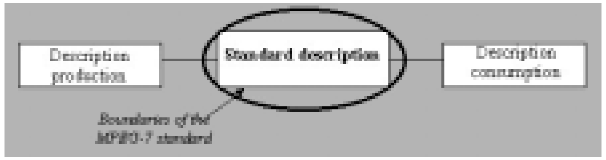


Combination of signal and semantic descriptors

- Semantic gap problem
- Separate search using each type of descriptor
  - Signal based: query by example
  - Symbol based: query by keywords, concepts, graphs
- Combination of both by association of confidence measures (weighted sum, product, minimum, ...)
- Crossing the gap during indexing and/or retrieval
  - Signal models of concepts
  - Concept recognition
- Relevance feedback
  - Implicit signal input by positive and negative examples

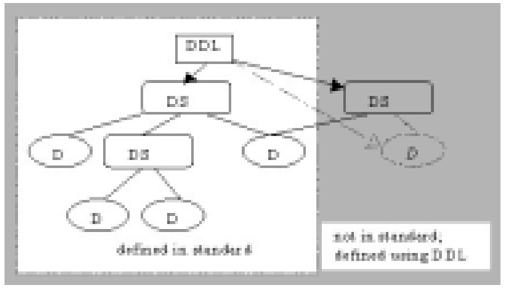
MPEG-7

- “Multimedia Content Description Interface”
- All levels, all application, all domains, ….



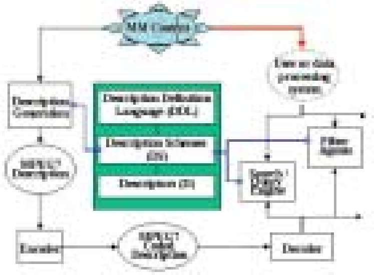
MPEG-7

- Descriptors (D), Description Schemes (DS) and Description Definition Language (DDL), plus ...



MPEG-7

- Descriptors (D), Description Schemes (DS) and Description Definition Language (DDL), plus ...



MPEG-7 parts

- *MPEG-7 Systems* - the tools that are needed to prepare MPEG-7 Descriptions for efficient transport and storage, and to allow synchronization between content en descriptions. Tools related to managing and protecting intellectual property
- *MPEG-7 Description Definition Language* - the language for defining new Description Schemes and perhaps eventually also for new Descriptors.
- *MPEG-7 Audio* – the Descriptors and Description Schemes dealing with (only) Audio descriptions
- *MPEG-7 Visual* – the Descriptors and Description Schemes dealing with (only) Visual descriptions
- *MPEG-7 Multimedia Description Schemes* - the Descriptors and Description Schemes dealing with generic features and multimedia descriptions
- *MPEG-7 Reference Software* - a software implementation of relevant parts of the MPEG-7 Standard
- *MPEG-7 Conformance* - guidelines and procedures for testing conformance of MPEG-7 implementations.

MPEG-7 Systems

- Tools that are needed to prepare MPEG-7 Descriptions for efficient transport and storage, and to allow synchronization between content and descriptions
- Tools related to managing and protecting intellectual property.
- Defines the terminal architecture and the normative interfaces

MPEG-7 Description Definition Language

- “... a language that allows the creation of new Description Schemes and, possibly, Descriptors. It also allows the extension and modification of existing Description Schemes.”
- XML Schema Language has been selected to provide the basis for the DDL. As a consequence of this decision, the DDL can be broken down into the following logical normative components:
  - The XML Schema structural language components;
  - The XML Schema datatype language components;
  - The MPEG-7 specific extensions.

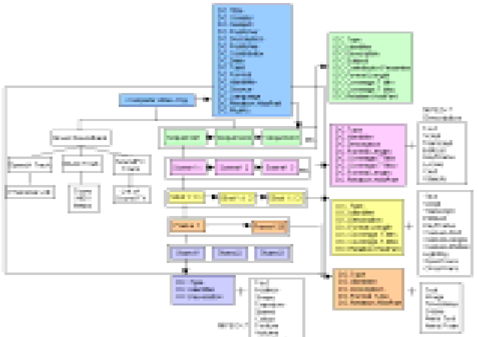
MPEG-7 Audio

- Audio description framework (which includes the scale tree and low-level descriptors),
- Sound effect description tools,
- Instrumental timbre description tools,
- Spoken content description,
- Uniform silence segment,
- Melodic descriptors to facilitate query-by-humming

MPEG-7 Visual

- Color
- Texture
- Shape
- Motion
- Localization
- Others

MPEG-7 / Dublin Core Index Structure



TREC Video Retrieval Evaluation

- A track / workshop designed to investigate content-based retrieval of digital video.
- Two-day workshop taking place just before TREC (17-18 nov. 2003).
- http://www-nlpir.nist.gov/projects/trecvid

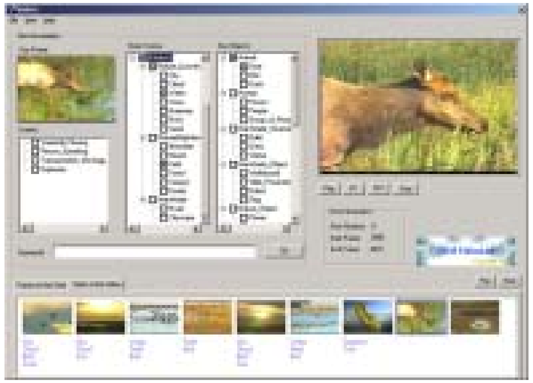
TREC Video Retrieval Tasks

- Shot Boundary Detection (2001-2003)
- Story Segmentation (2003)
- Feature Extraction (2002 : 10, 2003 :17)
- Search (2001-2003)
  - Manual (without relevance feedback)
  - Interactive (with relevance feedback)
- NOT a Spoken Document Retrieval task
  - Focus on visual queries.

TREC Video Retrieval Corpus

- 2001 : Open Video, NIST : 11 hours,
- 2002 : Open Video, Internet Archive, 73 hours (~44 GB), (very) old movies,
- 2003 : ABC news, CNN news, C-SPAN, 133 hours (~100 GB), restricted use, ~138,000 shots / key frames,
- MPEG-1 : 1.5 – 2.5 Mbit/s,
- From 2002 : donated speech transcription and donated features available,
- 2003 : closed captions available.

TREC Collaborative Video Annotation



TREC Video 2002 SBD results (Smeaton & Over, 2002)

Figure 1: Precision and recall for cuts

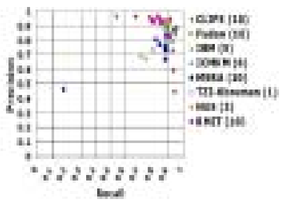
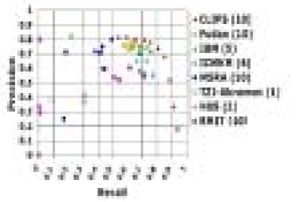


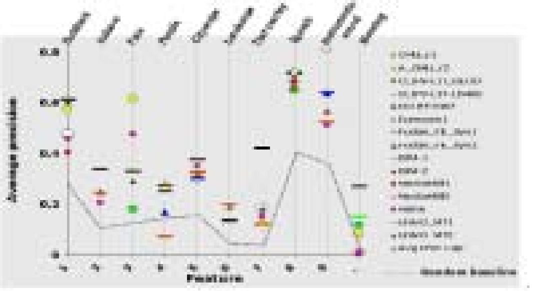
Figure 2: Precision and recall for gradual transitions



Equal false positive and negative:
Cuts: ~2-3 %   Gradual: ~20-30 %

TREC Video 2002 Feature results (Smeaton & Over, 2002)

Figure 4: Average precision by feature and run



TREC Video 2002 Topics (Smeaton & Over, 2002)

Table 2: Overview of topics

[illegible]

TREC Video 2002 Search results (Smeaton & Over, 2002)

Figure 6: Top 10 manual search runs

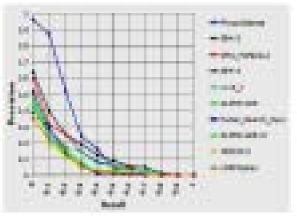
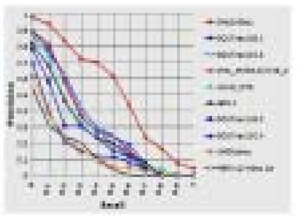


Figure 7: Top 10 interactive search runs



New in 2003: Story Segmentation

- Given the story boundary test collection, identify the story boundaries with their location (time) and type (miscellaneous or news) in the given video clip(s).
- The task is based on manual story boundary annotations made by LDC for the TDT-2 project.

Research Systems

- Audiosurf, LIMSI-CNRS (FR)
- AT&TV, AT&T Laboratories Cambridge (US)
- Informedia I et II Carnegie Mellon University (US)
- Físchlár (Dublin City University)
- CVSP, CLIPS-IMAG (FR)
- Cue Video, IBM Almaden Research Center (US)
- CWI Amsterdam (NL)
- Imperial College London (UK)
- Fudan Univ. (China)
- Prous Science (Esp)
- Microsoft Research Asia (China)
- ...

Research Systems

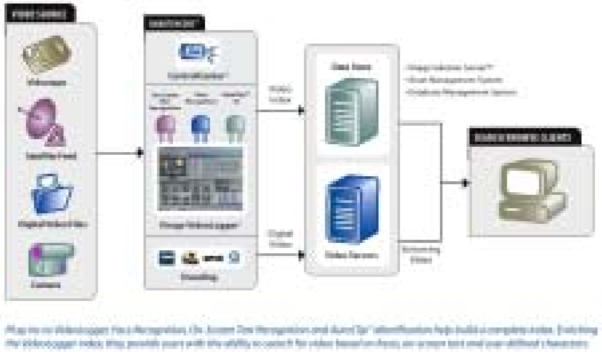
- Use of:
  - Automatic Speech Recognition
  - Automatic Text Recognition
  - Face detection and recognition
  - Feature extraction
  - Image similarity:
    - » Color
    - » Texture
  - Phonetic recognition plus word spotting
  - Orientation / scale filtering
  - Vectorial model
  - Relevance feedback

Commercial Systems

- Virage (Yahoo and Alta Vista)
- Convera (Excalibur)
- Easyglider
- LTU Technologies
- ...

- Complete offers
  - Indexing and Retrieval
  - Acquisition and streaming
  - Database managements
  - ...

Virage Complete Offer

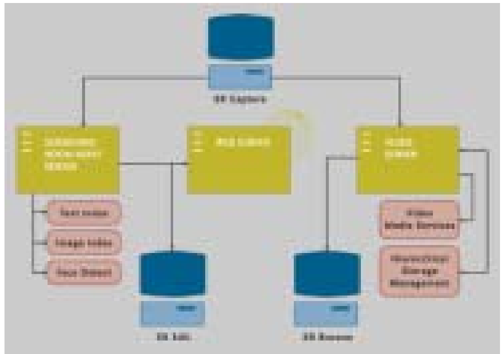


Virage Video Logger

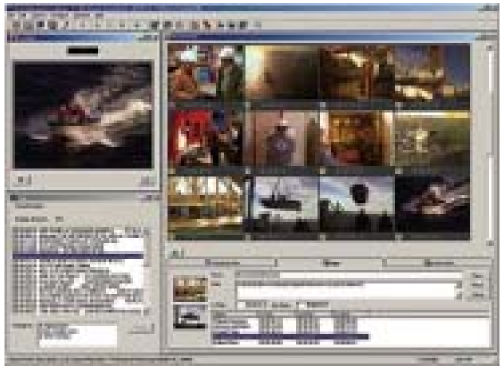
Virage Media Analysis Plug-ins

- Standard Audio Recognition: Identifies spoken words, speaker names and audio types and transforms them into searchable text in real time. Two audio recognition plug-ins are available, one for standard audio and one for advanced audio.
- Automatic Clip Recognition: Automatically segments video based on user-defined characters, words or symbols. The AutoClip? plug-in uses data from the text tracks (e.g. closed captioning, speech recognition, etc.) to automatically create in- and out-points based upon those parameters.
- Face Recognition: Checking against a library of user-defined faces, this plug-in recognizes people in the video frame and adds these names to the index.
- On-Screen Text Recognition: Recognizes text in the video frame, such as names in lower thirds, sports scores and product information.
- Advanced Audio Products:
  - Speech Recognition
  - Speaker Identification
  - Name Extraction
  - Story Recognition
- MediaSync?: Rapidly assembles and synchronizes streaming video with PowerPoint®. All the independent steps required to capture and encode video and synchronize it to PowerPoint can be orchestrated in on automated process. Providing transparent and seamless integration with PowerPoint and VirageVideoLogger, MediaSync is the easiest way to synchronize streaming video and slide presentations.

Convera (Excalibur) Screening Room



Convera (Excalibur) Screening Room



Web Video Retrieval Systems

	Text	Image	Video	Audio	M. Albright
http://www.netscape.com/	X	-	-	-	-
http://www.msn.com/	X	-	-	-	-
http://www.voila.fr/	X	-	-	-	-
http://www.google.com/	X	X	-	-	-
http://www.excite.com/	X	X	-	-	-
http://www.yahoo.com/	X	X	?	-	-
http://fr.altavista.com/video/default	X	X	X	X	72 (3/3)
http://multimedia.lycos.com/	X	-	X	-	2 (0/2)
http://www.alltheweb.com/	X	X	X	X	1
http://www.singingfish.com/	-	-	X	X	186 (3/3)
http://www.amazon.com/	X	-	X	X	1 (1/1)
http://speechbot.research.compaq.com/	-	-	-	X	179 (2/2)
http://www.ctr.columbia.edu/webseek/	-	X	-	X	0
http://www.tf1.fr/video/news/lesjt/	-	-	X	-	5 (1/5)
http://audiosurf.limsi.fr/ (*)	-	-	X	X	59 (1/3)

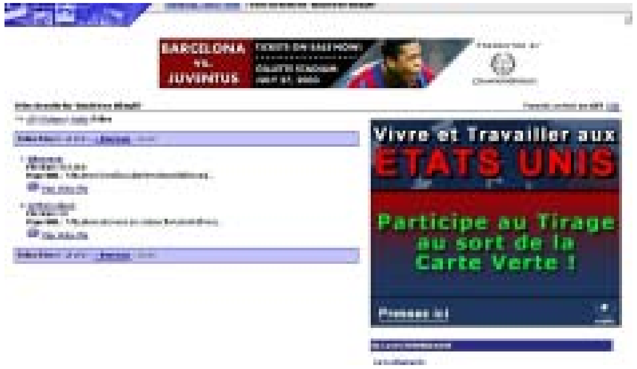
(*) password protected

http://fr.altavista.com/video/default

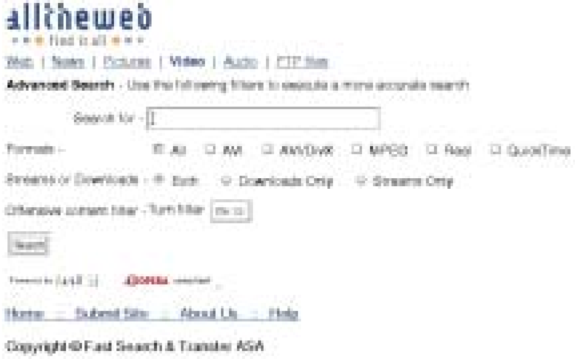
http://fr.altavista.com/video/default

http://multimedia.lycos.com/

http://multimedia.lycos.com/



http://www.alltheweb.com/


alltheweb
find it all
Web | News | Pictures | Video | Audio | FTP files
Search
Home :: Submit Site :: About Us :: Help

Advanced Search - Use the following filters to execute a more accurate search

Search for -

Formats - ☒ All □ AVI □ AVI/DivX □ MPEG □ Real □ QuickTime

Streams or Downloads - ● Both ○ Downloads Only ○ Streams Only

Offensive content filter - Turn filter On

Copyright © Fast Search & Transfer ASA

http://www.alltheweb.com/

alltheweb
Advanced Search :: Customize Preferences :: Submit Site :: Help
Web | News | Pictures | Video | Audio | FTP files
Home :: Submit Site :: About Us :: Help

Madeleine Albright

Title Length Size Format

Result Page: 1

http://www.singingfish.com/

Singingfish Advanced Audio Video Search

Madeleine Albright

http://www.singingfish.com/

Singingfish Advanced Audio Video Search

http://www.amazon.com/

Video

All Products

http://www.amazon.com/

Fiestaware

All Products

http://speechbot.research.compaq.com/

Search

All Topics

All Shows
AFRTS Radio News
American RadioWorks
Car Talk Radio Show
Dr. Toni Grant
Fresh Air
Geeks in Space

All dates

Relevance

http://speechbot.research.compaq.com/

http://disney.ctr.columbia.edu/webseek/

WebSEEk

A Content-Based Image and Video Search and Catalog Tool for the Web

(press here to Browse all subjects)

Animals
Architecture
Art
Astronomy
Cats
Celebrities
Dogs
Food
Horror
Humour
Movies
Music
Nature
Sports
Transportation
Travel

Image/Video Topic

Search

http://www.tf1.fr/video/news/lesjt/

tf1.fr

http://www.tf1.fr/video/news/lesjt/

tf1.fr

VIDEO

http://audiosurf.limsi.fr/

LIMSI-CNRS

http://audiosurf.limsi.fr/

Conclusion

- Models for content representation available
  - Both at signal and semantic levels
  - How to fill them ? (indexing)
  - How to exploit them ? (search)
  - Still room for improvements
- Semantic gap still largely open:
  - Signal, low level, automatic, not so useful
  - Semantic, high level, manual, slow and costly
  - Two significant exceptions:
    - People detection and recognition
    - Speech transcription
  - Need for “AI style” content understanding techniques
  - Need to manage general and domain specific knowledge bases, world representations.