ERMITES 2008

Multimedia Indexing

Prof. Bernard Merialdo

Institut Eurecom

merialdo@eurecom.fr

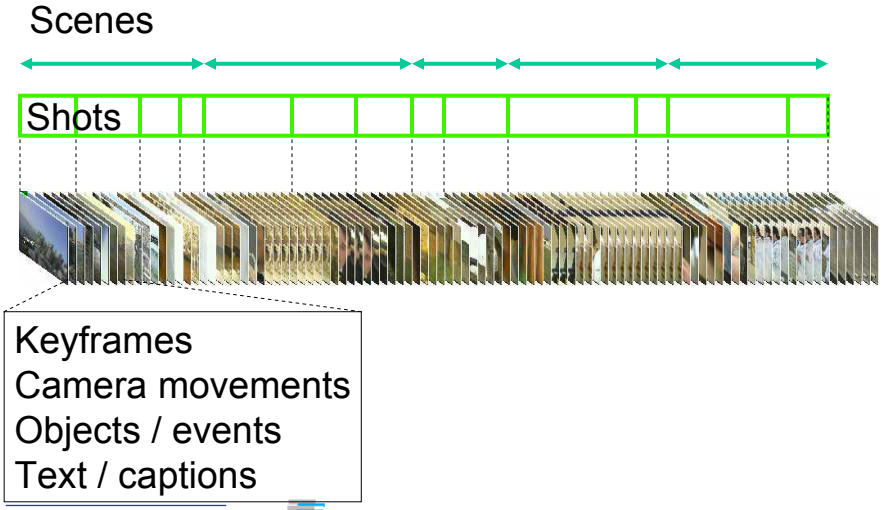
EURECOM Sophia Antipolis

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Contents

- TrecVid
- Shot Segmentation
- Semantic Classification
- Video Search
- Summarization
- MPEG-7

Video Indexing



TrecVid

- Evaluation campaign organized by NIST (National Institute of Standards, USA)
- Purpose: compare video retrieval algorithms on same data and tasks
- Started in 2001 as a track of TREC
- Independant campaign from 2003
- Participants: 12 in 2001, 54 in 2007

TrecVid Data

Year	Hours of video (training/test)	Type
2001	11	NIST videos
2002	73	Internet Open Archive
2003	66/67	TV News (ABC, CNN, CSPAN)
2004	0/70	TV News (ABC, CNN, CSPAN)
2005	85/85	TV News (+arabic, chinese)
2006	0/158	TV News (+arabic, chinese)
	50	BBC Rushes
2007	50/50	Sound and Vision (dutch)
	18/17	BBC Rushes
2008	100/100	Sound and Vision (dutch)
	35/18	BBC Rushes
	200	Surveillance

TrecVid Tasks

- Shot Boundary Determination 2001-2007
- Search 2001-2008
- High-Level Feature Extraction 2003-2008
- Stories 2003-2004
- BBC Rushes 2005-2008
- Camera motion 2006
- Surveillance 2008
- Copy detection 2008

Shot Segmentation

- A shot is a continuous take from one camera
- The transition from one shot to the next can be a hard cut or a gradual transition
- Hard cuts can generally be easily detected:

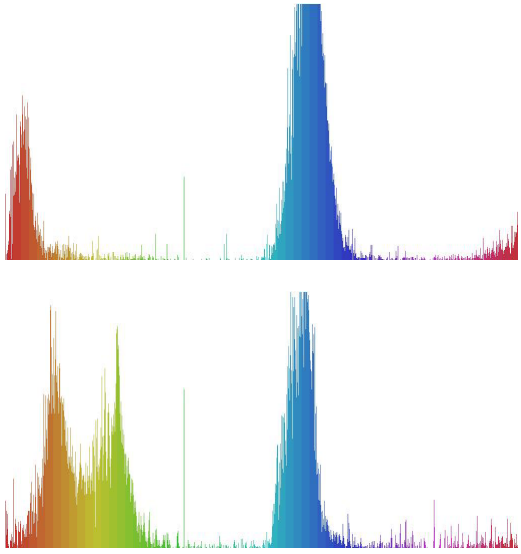
- Gradual transitions span over several frames
- There are many types of gradual transitions based on different visual effects

Shot Segmentation

- Dissolve

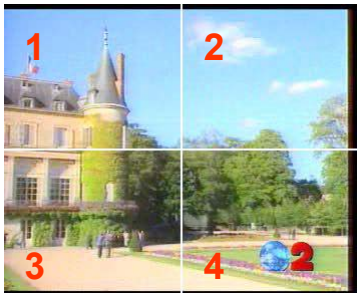
  - Special case: fade-in, fade-out
- Wipe

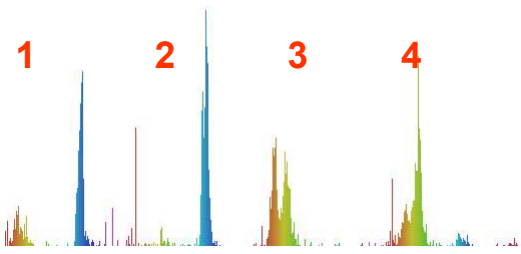
Color Histogram: per keyframe



Color Histogram: region-based

- Split the image into regions, concatenate the region histograms





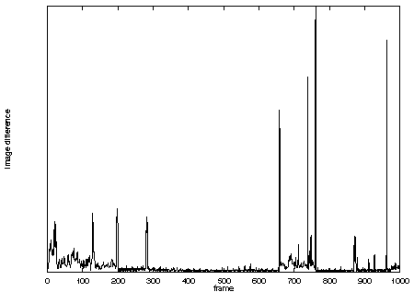
Cut Detection: hard cuts

- Basic idea:
  - Measure distance $d(I_t, I_{t+1})$ between consecutive frames
  - Detect cut if distance is greater than threshold:

    $d(I_t, I_{t+1}) \geq \theta$
- Common distance: color histogram

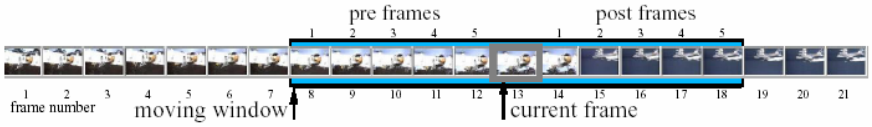
$$d(I_t, I_{t+1}) = \sum_{c \in Colors} |h_t(c) - h_{t+1}(c)|$$

  - Depends on color space
  - Robust to object movements
  - Efficient for hard cuts
  - Poor for gradual transitions



Cut Detection: Gradual Transitions

- Sliding window:



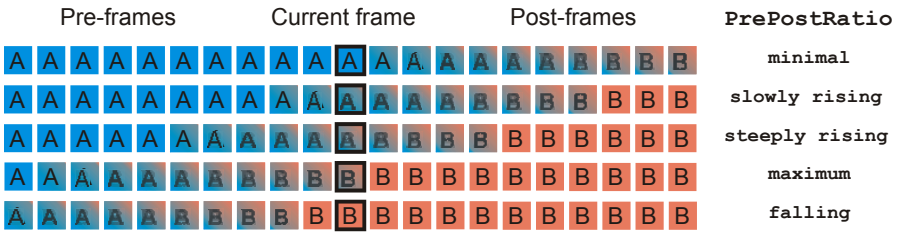
  - Compare pre- and post- frames with current frame f_c
  - Compute PrePostRatio:

$$PrePostRatio = \frac{\sum_{f \in PreFrames} d(f, f_c)}{\sum_{f \in PostFrames} d(f, f_c)}$$

  - Peak of PrePostRatio = end of gradual transition

Cut Detection: Gradual Transitions

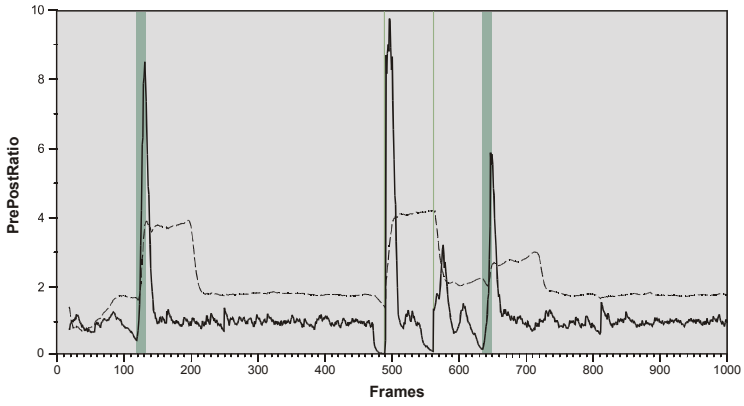
- Dissolve between shot A and shot B:



- PrePostRatio is usually minimal at the beginning of a gradual transition and rises up to a maximum at the end of the transition

Cut Detection: Gradual Transitions

- Example of PrePostRatio curve



- two short gradual transitions and two cuts

Cut detection: difficult cases

- Similar environment

  - Change in camera position
  - Same color ambiance
  - Cut is difficult to detect

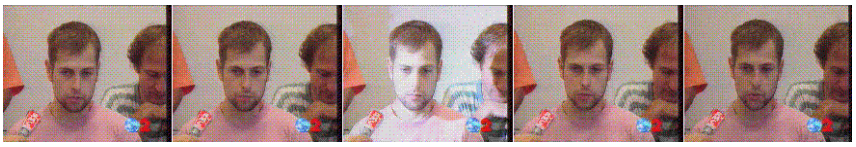
Cut detection: difficult cases

- Fast movement of large object

  - Can be confused with wipe
  - Shot can be over-segmented

Cut detection: difficult cases

- Sudden change in illumination

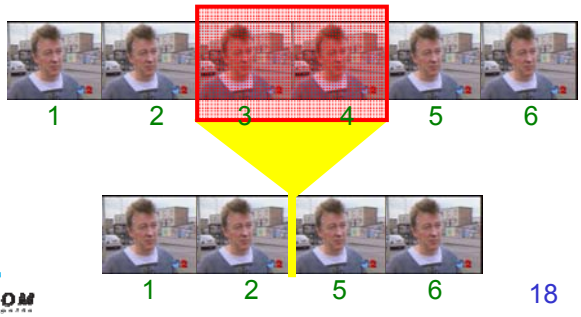


  - Sudden modification of colors
  - Also the case in explosions, etc…
  - Shot can be over-segmented

Cut detection: ambiguous cases

- Inserts

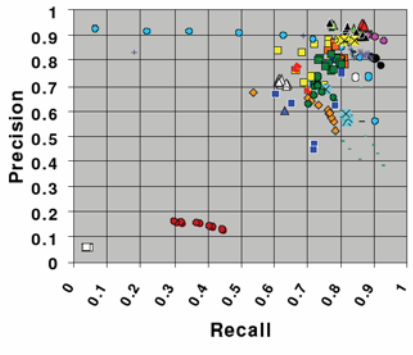
- Interview edit



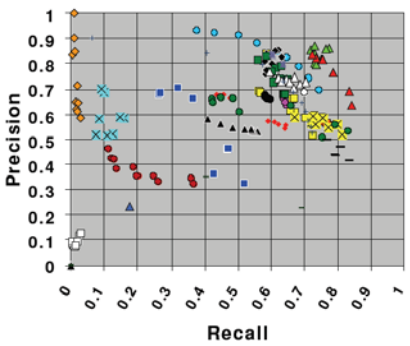
TrecVid: Shot Boundary Determination

- 2006 Results:
  - 13 news videos, 3785 transitions

Cuts



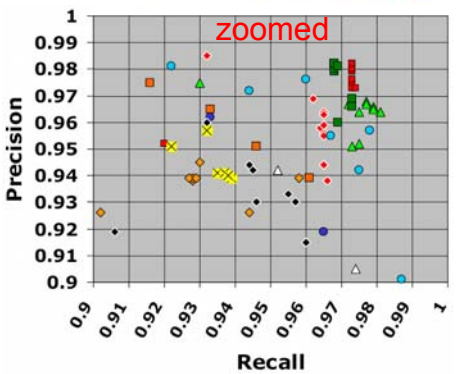
Gradual transitions



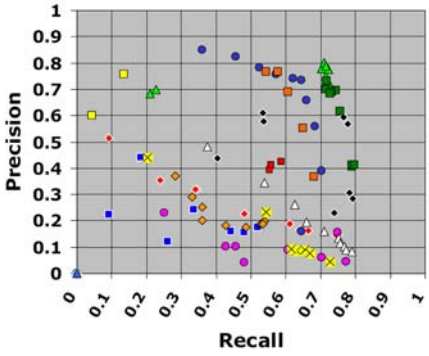
TrecVid: Shot Boundary Determination

- 2007 Results:
  - 17 news videos, 2463 transitions

Cuts



Gradual transitions



TrecVid: High Level Feature Extraction

- Task: decide if a shot contains a given concept or not
- Objective: build generic concept detectors
- Method:
  - Assume high level feature is binary (contains or not contains)
  - Rank shots with concept by confidence
  - Return best 2000 shots for each feature
  - Manual assessment by NIST (on 20 features)
  - Compute Mean Average Precision (MAP)

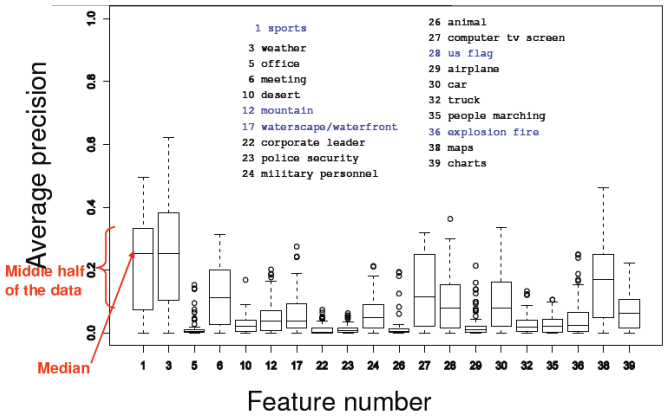
TrecVid: High Level Feature Extraction

- Goal: decide if a shot contains a concept or not
- 39 concepts:

Sports	Sky	Computer_TV-screen
Entertainment	Snow	Flag-US
Weather	Urban	Airplane
Court	Waterscape_Waterfront	Car
Office	Crowd	Bus
Meeting	Face	Truck
Studio	Person	Boat_Ship
Outdoor	Government-Leader	Walking_Running
Building	Corporate-Leader	People-Marching
Desert	Police_Security	Explosion_Fire
Vegetation	Military	Natural-Disaster
Mountain	Prisoner	Maps
Road	Animal	Charts

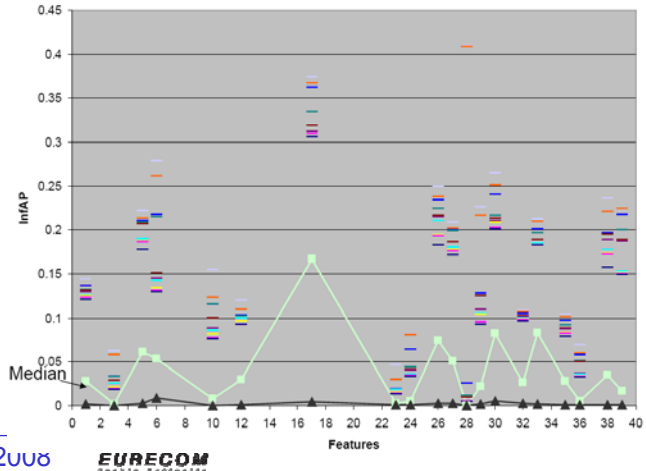
TrecVid: High Level Feature Extraction

- 2006: 20 concepts evaluated
- Overall results:



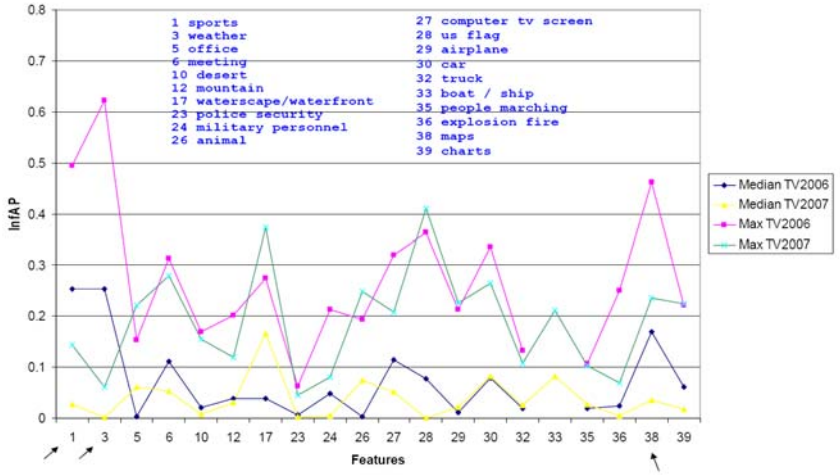
TrecVid: High Level Feature Extraction

- 2007: 20 concepts evaluated
- Overall results:



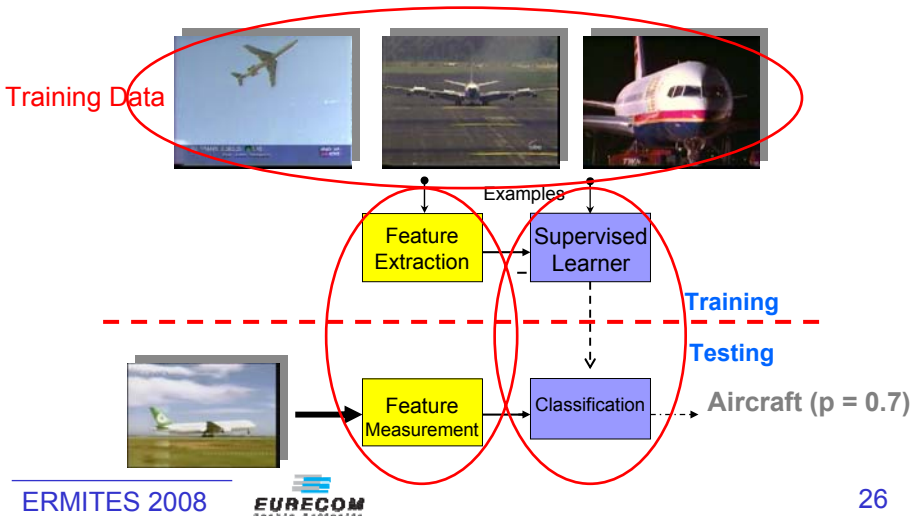
TrecVid: High Level Feature Extraction

- 2007 vs 2006



TrecVid: High Level Feature Extraction

- Generic concept detector:



TrecVid: High Level Feature Extraction

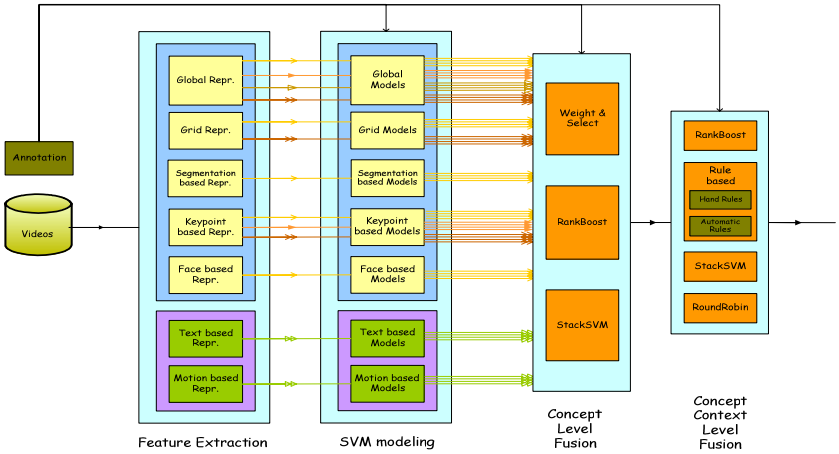
- Training data
- 2005 Collaborative annotation by TREC participants

TrecVid: High Level Feature Extraction

- Training data
- LSCOM : Large Scale Concept Ontology for Multimedia
  - Project by Columbia U, IBM, CMU
  - 856 concepts designed for TV News
    - Events, locations, people, programs…
  - Manual annotation of TV2005 videos for 449 concepts (61901 shots)
  - Useful for large scale experiment and cross-concept correlations

TrecVid: High Level Feature Extraction

- Tsinghua system (best TV2006 performer)



TrecVid: High Level Feature Extraction

- Image features
  - Color histogram
  - Texture
  - Edge
- Audio features
  - FFT
  - MFCC
- Motion features
  - Kinetic energy
  - Optical flow
- Detector features
  - Face detection
  - Video-OCR detection

TrecVid : CMU Detector features

- Face detector
  - Detecting faces in the images

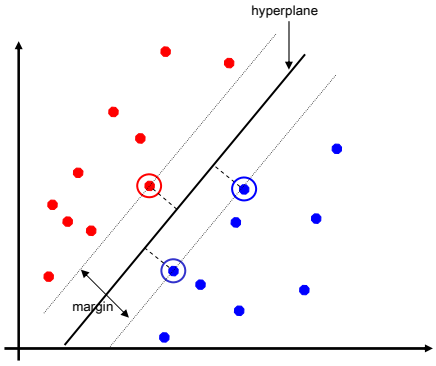
- VOCR detector
  - Detecting and recognizing VOCR

TrecVid: High Level Feature Extraction

- Classifiers:
  - SVM
  - Neural Networks
  - K-NN
  - Adaboost
  - Etc…

TrecVid : SVM Classifier

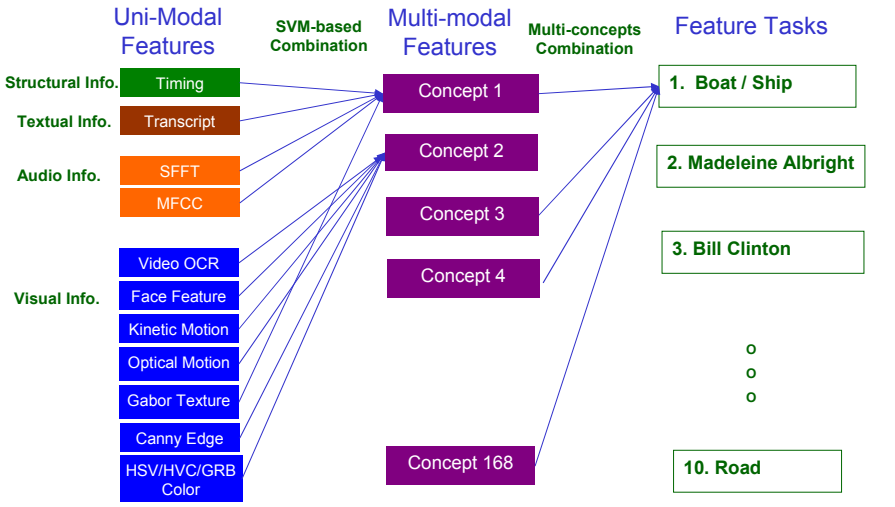
- Binary classifier
- Constructed from training data
- Linear separator with highest margin in space with Kernel distance



TrecVid: High Level Feature Extraction

- Many features, many classifiers, but not one better than all others
- Need to combine all information to build best system: information fusion
- Several approaches:
  - Early fusion of feature vectors
  - Late Fusion of classifier decisions

TrecVid : CMU Architecture

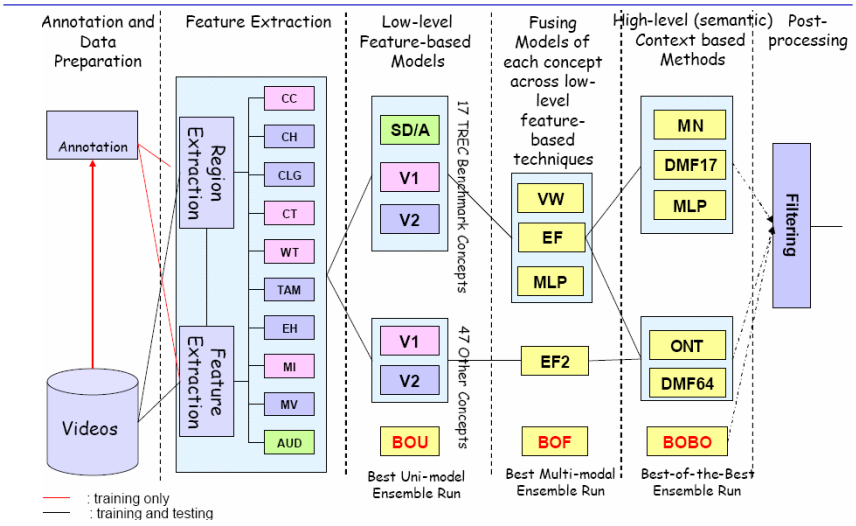


TrecVid : CMU Multi-concepts Combination

- Bayesian Networks from 168 common annotation concepts
- Combine 4 most related concepts with target concept

Boat/Ship	Boat, Water_Body, Sky, Cloud
Train	Car_Crash, Man_Made_scene, Smoke, Road
Beach	Sky, Water_Body, Nature_Non-Vegetation, Cloud
Basket Scored	Crowd, People, Running, Non-Studio_Setting
Airplane Takeoff	Airplane, Sky, Smoke, Space_Vehicle_Launch
People Walking/running	Walking, Running, People, Person
Physical violence	Gun_Shot, Building, Gun, Explosion
Road	Car, Road_Traffic, Truck, Vehicle_Noise

TRECVID: IBM Pipeline



TrecVid: Search Task

- Goal: find the shots satisfying a query
- Queries are defined by text + sample keyframes + sample shots
- 2006 Topic examples:
  - Topic 173: Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)
  - Topic 174: Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible
  - Topic 175: Find shots with one or more people leaving or entering a vehicle
  - …
  - Topic 195: Find shots of one or more soccer goalposts
  - Topic 196: Find shots of scenes with snow

TrecVid: Search Task

Deployment

Find shots of one or more helicopters in flight.

Find shots of a hockey rink with at least one of the nets fully visible from some point of view.

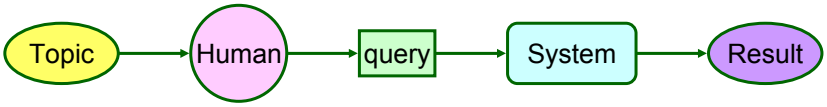
Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

Find shots of a group including at least four people dressed in suits, seated, and with at least one flag.

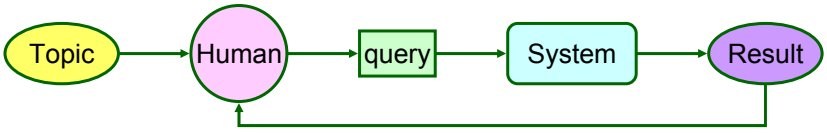
TrecVid: Search Task

- 3 types of experiments:
  - Automatic:

  - Human-assisted:

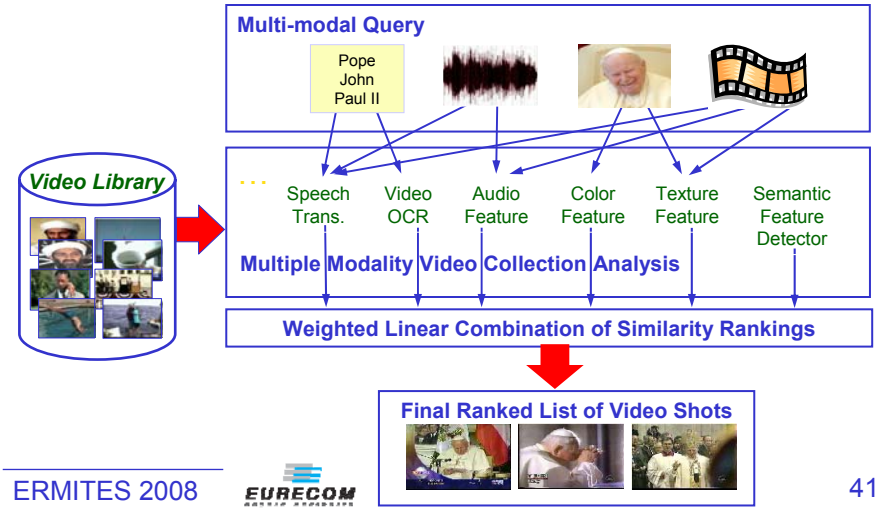


  - Interactive:



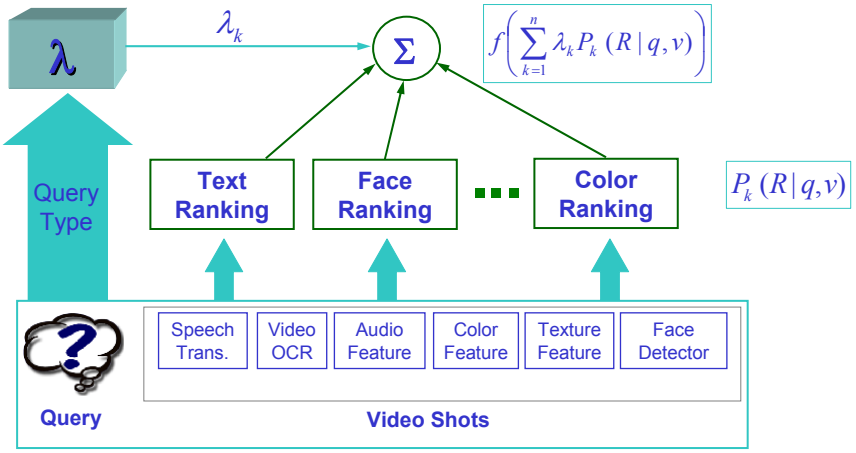
TrecVid: Search Task Example

- CMU Automatic Search



TrecVid: Search Task Example

- Query-type dependent weights

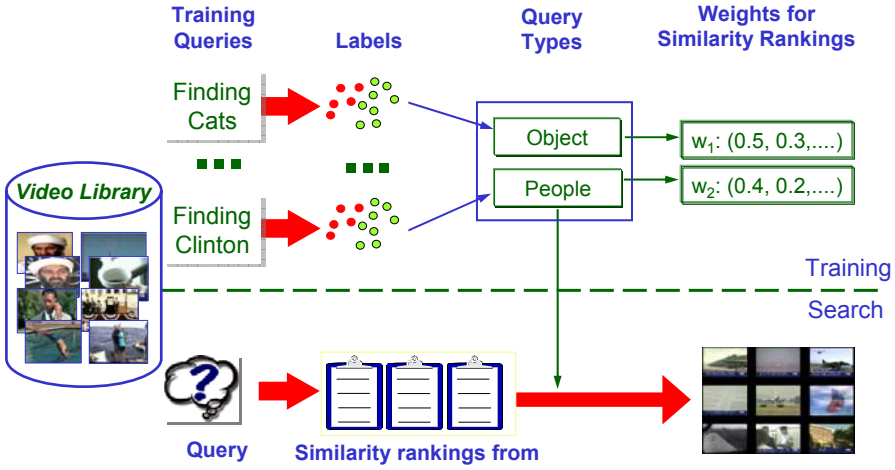


TrecVid: Search Task Example

- Possible query types PEOS:
  - Named person queries (P-queries)
    - "Find shots of Yasser Arafat"
      "Find shots of Ronald Reagan speaking"
  - Named entity queries (E-queries)
    - "Find shots of the Statue of Liberty"
      "Find shots of the Mercedes logo"
  - General object queries (O-queries)
    - "Find shots of snow-covered mountains"
      "Find shots of one or more cats"
  - Scene queries (S-queries)
    - "Find shots of roads with lots of vehicles"
      "Find shots of people spending leisure time on the beach"

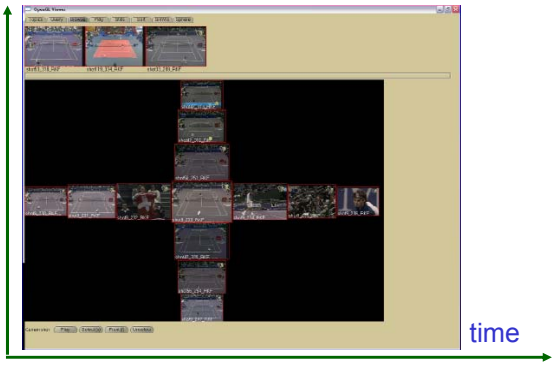
TrecVid: Search Task Example

- Learning Query-type dependent weights



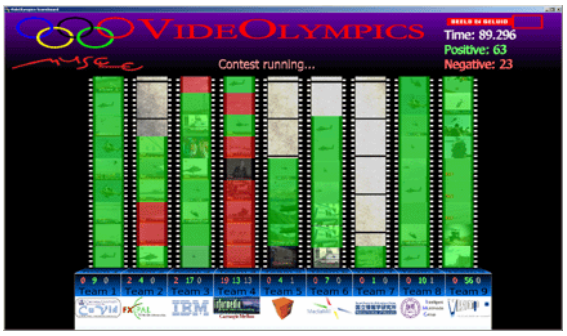
TrecVid: Search Task Example

- 2005 MediaMill Interactive Search (Amsterdam)
- Use 101 concept detectors
- Cross Browser to explore multimedia space



VideOlympics

- Organized at CIVR 2007 (and 2008)
  - Parallel public use of search systems
  - Comparative live panel of shots found



TRECVID BBC Rushes summarization task

- Rushes from BBC archive
  - Unedited material from five dramatic series
  - 18 hours (43 videos) for development
  - 17 hours (42 videos) for testing

* http://www-nlpir.nist.gov/projects/trecvid/

TRECVID BBC Rushes summarization task

- Summarization task
  - Create an MPEG-1 summary of each file
  - **Eliminate redundancy**
  - Maximize viewers' efficiency at recognizing objects & events as quickly as possible
  - Interaction limited to simple playback with optional pauses

TRECVID BBC Rushes summarization task

- Ground truth : human annotation of visible topics
- Sample for MRS044500 :
  - 2 men in dark suits walk past Ford truck to building entrance
  - 2 men in dark suits enter building
  - person in brown coat opens rear end car and removes wheelchair (seen from front of car)
  - woman walks around car to passenger window (seen from rear end of car)
  - close up of man in passenger seat (seen from front of car)
  - woman in brown coat removes wheelchair and brings it round to the passenger door (seen from front of car)
  - man in beige suit appears (seen from front of car)
  - man in beige suit opens car door (seen from front of car)
  - woman in brown jacket undoes man in car's seatbelt (seen from front of car)
  - woman in brown jacket helps passenger into wheelchair (seen from front of car)
  - …

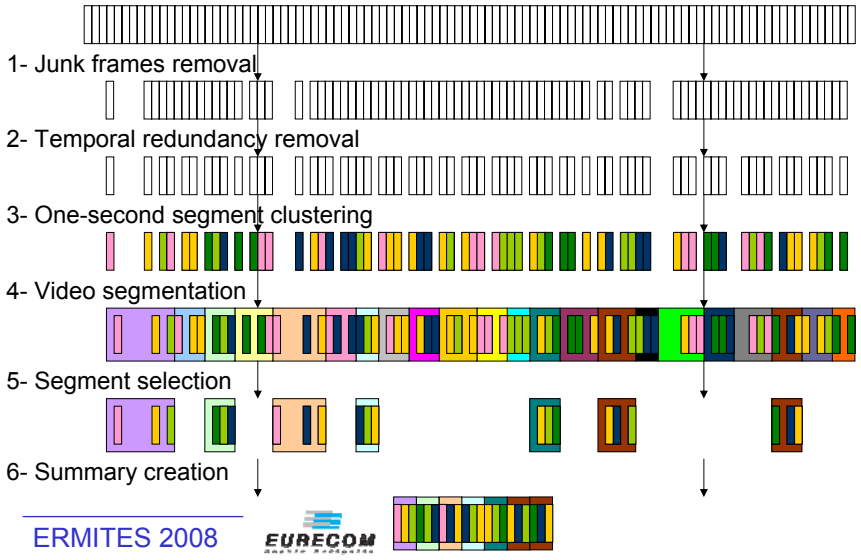
TRECVID BBC Rushes summarization task

- Evaluation :
  - For each video, 12 topics are selected at random
  - Evaluator watches summary (with pauses)
  - Checks which topics (out of 12) are present
  - Various measures:
    - Fraction of (12 items of) ground truth found
    - Ease of use
    - Amount of near-redundancy
    - Assessment time to judge included ground truth
    - Summary duration
    - Summary creation compute time
    - Number/duration of pauses in assessment of included segments
    - Feedback on assessment software, procedure, experience

TRECVID BBC Rushes summarization task

- 2005: no task
- 2006: organize, no evaluation
- 2007: summarize, evaluate
  - List of topics and events built for ground truth
  - 4% summary is built for each video
  - Evaluator watches summary and counts topics present
- 2008: summarize, evaluate
  - 2% summary is built for each video
  - Evaluator watches summary and counts topics present

Eurecom Summarization System



Eurecom Irrelevant Frames Removal

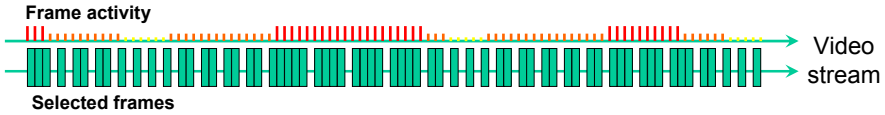
- **Test pattern frame modeling**
  - Training set: 2452 test pattern frames
  - Feature vector: HSV Color histogram
  - Euclidian distance
  - Model: mean vector + threshold
- **Uniform color frame modeling**
  - Color pixel distribution entropy

$$E = -\sum_{c} p(c) Log\, p(c)$$

  - Detection: E < threshold

Eurecom Irrelevant Frames Removal

- Temporal Redundancy
  - Dynamic acceleration proportional to the motion activity



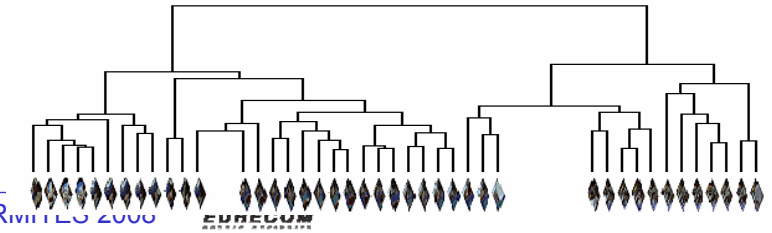
  - Frame activity: percentage of modified pixels *activ(f)*
  - Jump threshold:

$$jump(v) = \frac{\sum_{f \in v} activ(f)}{\text{VideoLength}} \times \text{MeanAcceleration}$$

  - After frame f_t, select f_k such that: $\sum_{i=t+1}^{k} activ(f_i) \geq jump(v)$

Eurecom Segment Selection

- Bottom-up Hierarchical clustering
  - Video segmentation in one-second segments
  - Feature vector: HSV histogram
    - Initially : one cluster per one-second segment
    - Iterations: merge closest clusters
      - Distance between two segments : Euclidian distance
      - Distance between two clusters : average distance across all possible pairs of segments



Eurecom Segment Selection

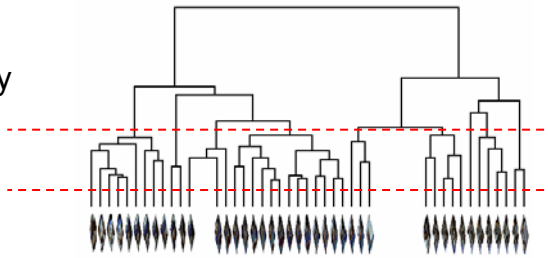
- Visual interest of a cluster $V_{int}(C)$
  - Each cluster represents a part of the visual space
  - Initially, all clusters are equally important:
  - $V_{int}(C)=1$
  - Improvement: we can analyse the content of a cluster based on segment indicators:
    - The appearance of people : *face(s)*
    - The activity : *activity(s)*
    - The color diversity : *entropy(s)*
  - Improved definition of visual interest

$$V_{\text{int}}(C)=\alpha*\frac{\sum_{s\in c}(face(s)*W_{\text{face}}+activity(s)*W_{\text{activity}}+entropy(s)*W_{\text{entropy}})}{W_{\text{face}}+W_{\text{activity}}+W_{\text{entropy}}}+(1-\alpha)$$

Eurecom Segment Selection

- We want to avoid selecting «similar» segments
- Similarity level depends on video:
  - Sometimes video content is very diverse
  - Sometimes video is quite static
- Selecting different levels of the hierarchy allows various levels of similarities

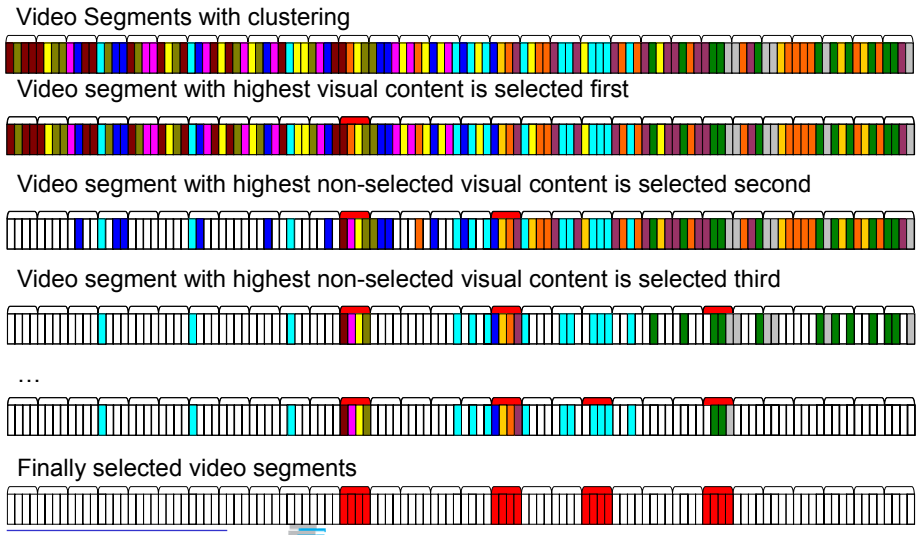
Coarser similarity
Finer similarity



Eurecom Segment Selection

- A Video Segment is a sequence of one-second segments of predefined length
- For a given hierarchy level, we iteratively select the most important and non-redundant Video Segments
  - The importance of a Video Segment is the sum of the visual interest of the (non-selected) clusters that it contains
  - We select Video Segments until all clusters have been selected
- We adjust the level so that the total duration of selected Video Segments is closest to the expected duration

Eurecom Segment Selection

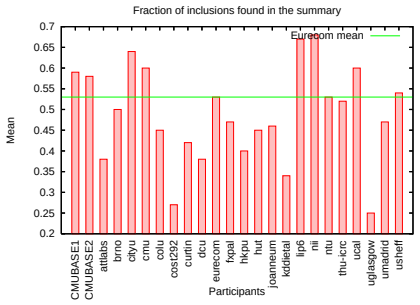


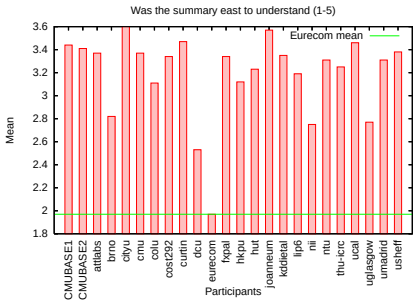
Eurecom Presentation

- 2006 Split-scren display
  - Maximize information displayed by time unit
  - Group shots by 4
- 2007
  - Main frame
  - Timeline
  - Icons for keyframe summary

TRECVID Summarization Evaluation

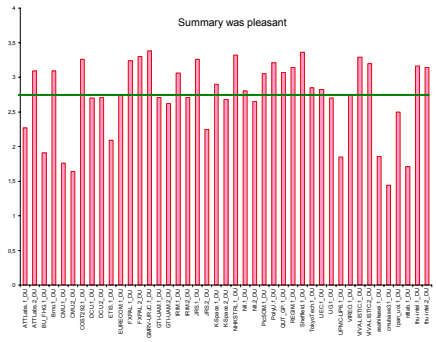
- 2007 Comparative results:
  - Good on inclusions
  - Low on usability





TRECVID Summarization Evaluation

- 2007 Comparative results:
  - Not so good on inclusions
  - Fair on usability
  - Surprizing results for baseline



Automating TRECVID Evaluation

- TRECVID 2007 : manual evaluation
  - Random list of 12 topics from groundtruth is selected
  - Assessor views the summary, checks topics that are present
  - Performance indicators are computed
- Problem:
  - Manual evaluation is difficult to implement and to reproduce
  - Difficult to make lots of experiments (tune parameters)
  - Difficult to define « groundtruth summary »
- Our solution:
  - We developed an automatic evaluation method
  - Idea:
    - The appearances of each topics are time-stamped
    - If one second of the topic is in the summary, the topic is considered to be found
  - Strong correlation between automatic and manual method
  - Admissible for comparing results

Automating TRECVID Evaluation

- 7 Test Videos
- Manual annotation:
  - Topics from TrecVid
  - Timestamps manually added:

T_1 $t_{11} - t_{12}$ $t_{13} - t_{14}$ $t_{15} - t_{16}$

T_2 $t_{21} - t_{22}$ $t_{23} - t_{24} \ldots$

- Performance measures:

$$\text{Recall} = \frac{\text{number of topics found}}{\text{number of topics}}$$

$$\text{Precision} = \frac{\text{number of topics found}}{\text{number of segments selected}}$$

Automating TRECVID Evaluation

- For each topic, build a feature vector from ground truth and summary, and try to predict assessment

GT vs video

Topic	video	#seq	min	max	mean	activity	entropy
1	MRS044500	4	95	1524	805	0.322	3.324
2	MRS044500	3	155	210	180	0.317	4.905
…	…	…	…	…	…	…	…

GT vs Summary

#seq	min	max	mean
4	7	103	56
1	0	26	8
…	…	…	…

Modeling and Prediction

Assessment

A_k
Yes
No
…

- Use various Machine Learning techniques

Automating TRECVID Evaluation

- Compare prediction X with real assessment Y
- Pearson correlation coefficient (reflects the degree of linear relationship between two variables).

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{V(X)*V(Y)}}$$

- Results

Id assessor	A1	A2	A3	A4	A5	A6	A7	AD Tree	Bayes Net	Stump
Pearson	0.738	0.532	0.587	0.607	0.652	0.658	0.637	0.69	0.77	0.70

MPEG-7

Multimedia Content Description Interface

EURECOM Sophia Antipolis

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

MPEG History

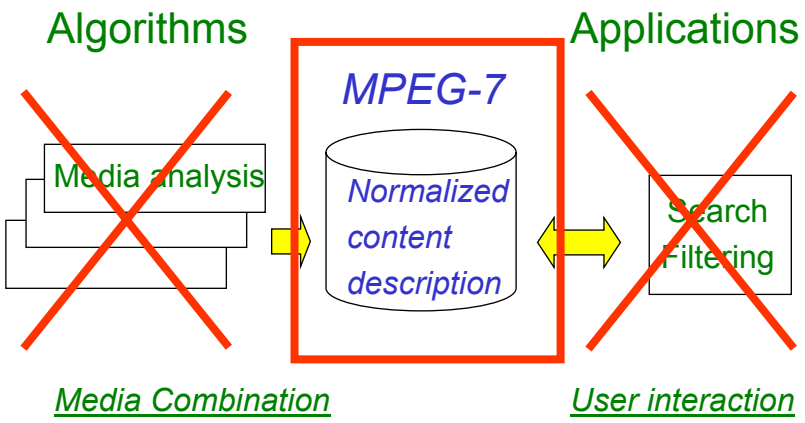
- MPEG = Moving Picture Experts Group
  - Started in 1988, Leonardo Chiariglione

- MPEG-1: Interactive CD and MP3 1992
- MPEG-2: DTV, STB, DVD 1994
- MPEG-4: Web and Mobility 1998-1999
- MPEG-7: Multimedia Content Description Interface 2001
- MPEG-21: Multimedia Framework ---

MPEG-7 Objective

- « Standardize content-based description for various types of audio-visual information, allowing quick and efficient content identification, and addressing a large range of applications »
- MPEG-7:
  - Information about the content
  - The bits about the bits
  - Metadata

MPEG-7 Scope

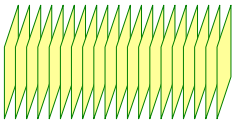


MPEG-7 Components

- MPEG-7 Systems
- MPEG-7 Description Definition Language
- MPEG-7 Visual
- MPEG-7 Audio
- MPEG-7 Multimedia DSs
- MPEG-7 Reference Software
- MPEG-7 Conformance

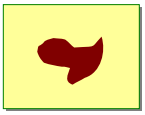
Low level Audio Visual descriptors

Video segments



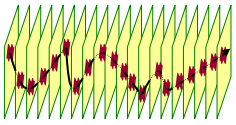
- Color
- Camera motion
- Motion activity
- Mosaic

Still regions



- Color
- Shape
- Position
- Texture

Moving regions

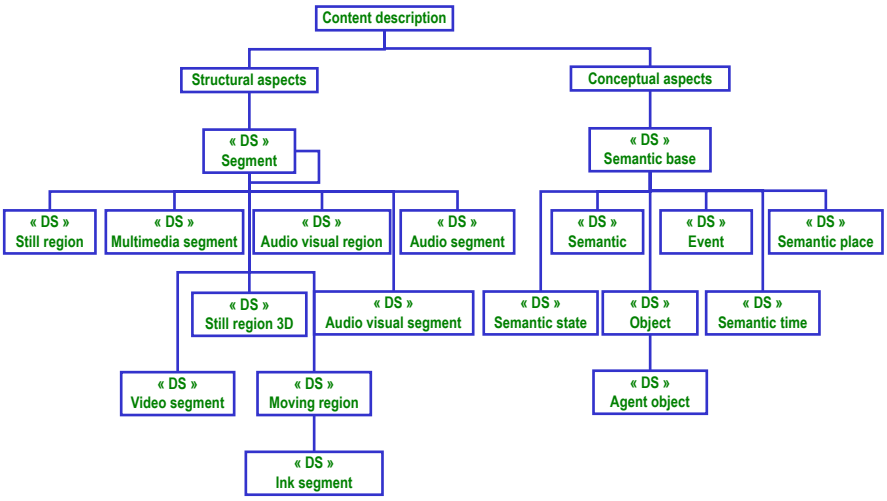


- Color
- Motion trajectory
- Parametric motion
- Spatio-temporal shape

Audio segments

- Spoken content
- Spectral characterization
- Music: timbre, melody

Multimedia DS: Content Description



MPEG-7: Application Areas

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