ERMITES 2007

Multimedia Indexing

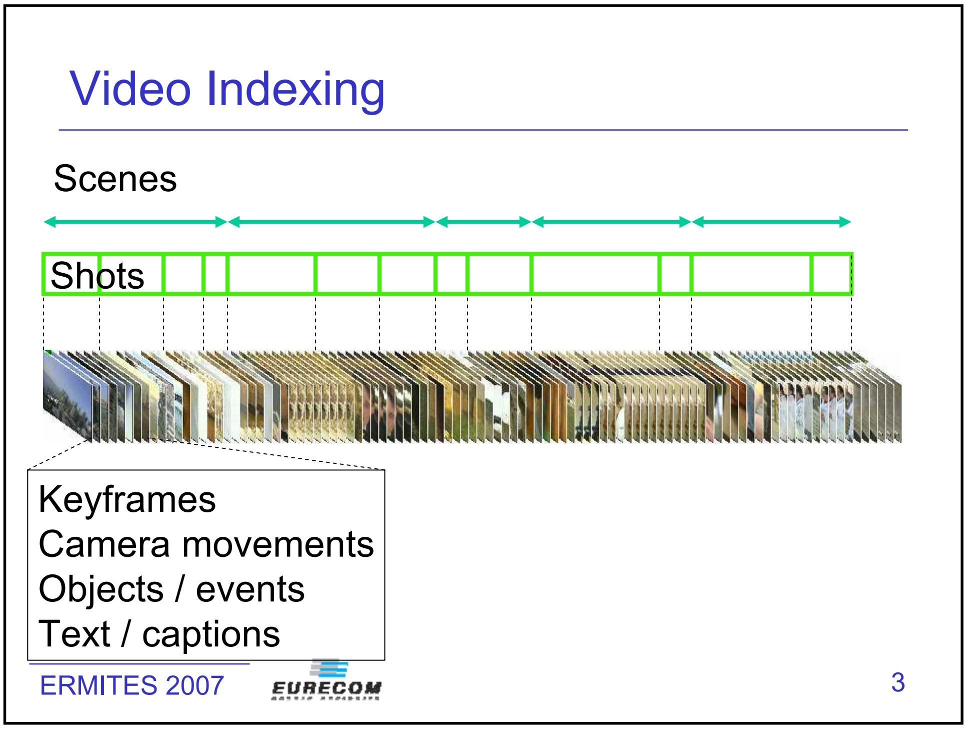
Prof. Bernard Merialdo

Institut Eurecom

merialdo@eurecom.fr

Contents

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

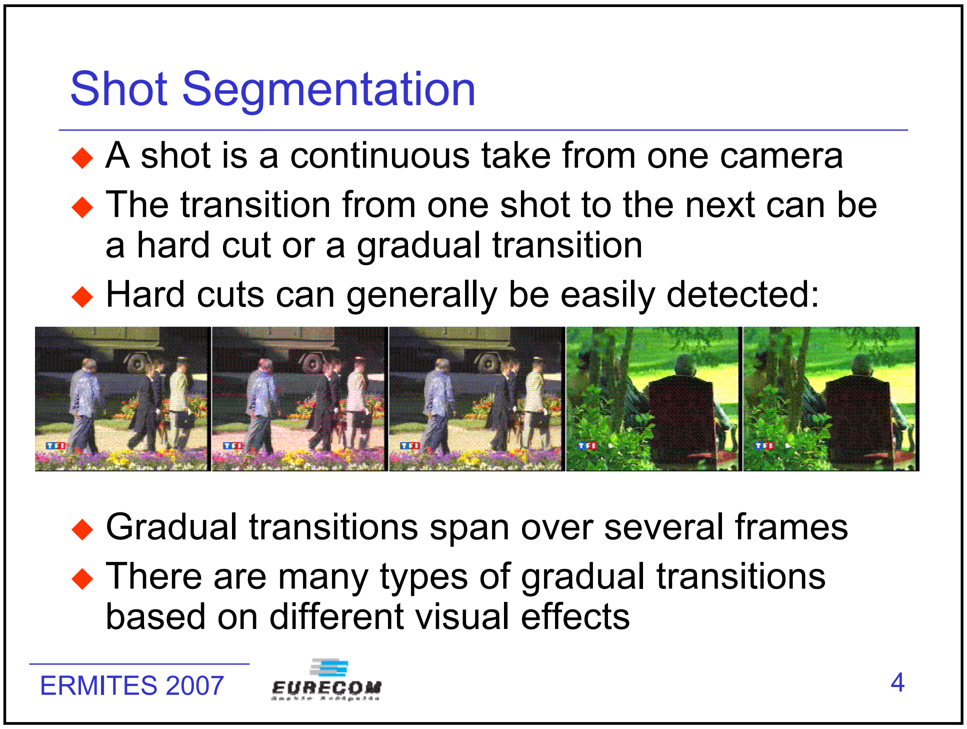


Video Indexing

Scenes

Shots

Keyframes
Camera movements
Objects / events
Text / captions

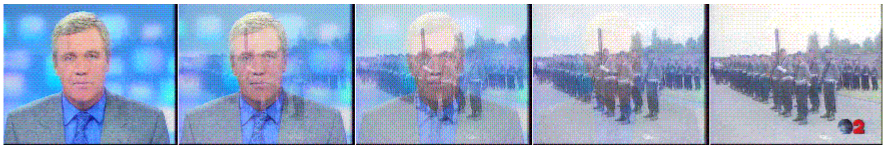


Shot Segmentation

- 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

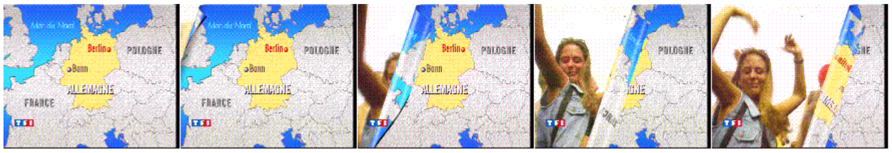
Shot Segmentation

- Dissolve

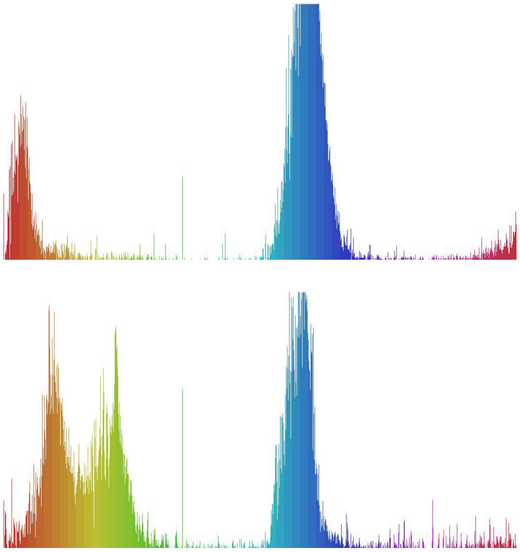


  - Special case: fade-in, fade-out

- Wipe

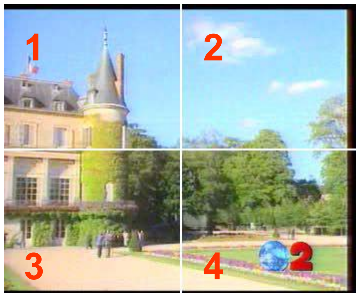


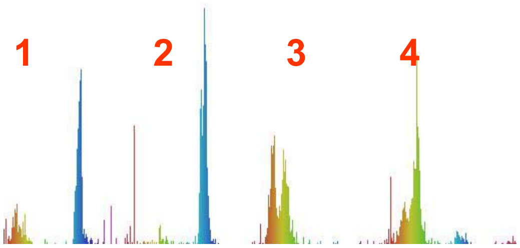
Color Histogram: per keyframe



Color Histogram: region-based

- Split the image into regions, concatenate the region histograms



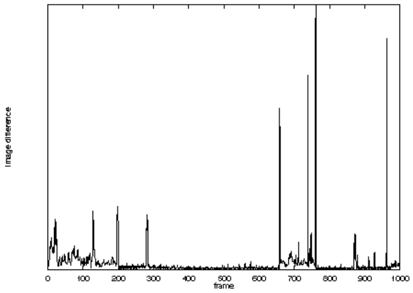


Cut Detection: hard cuts

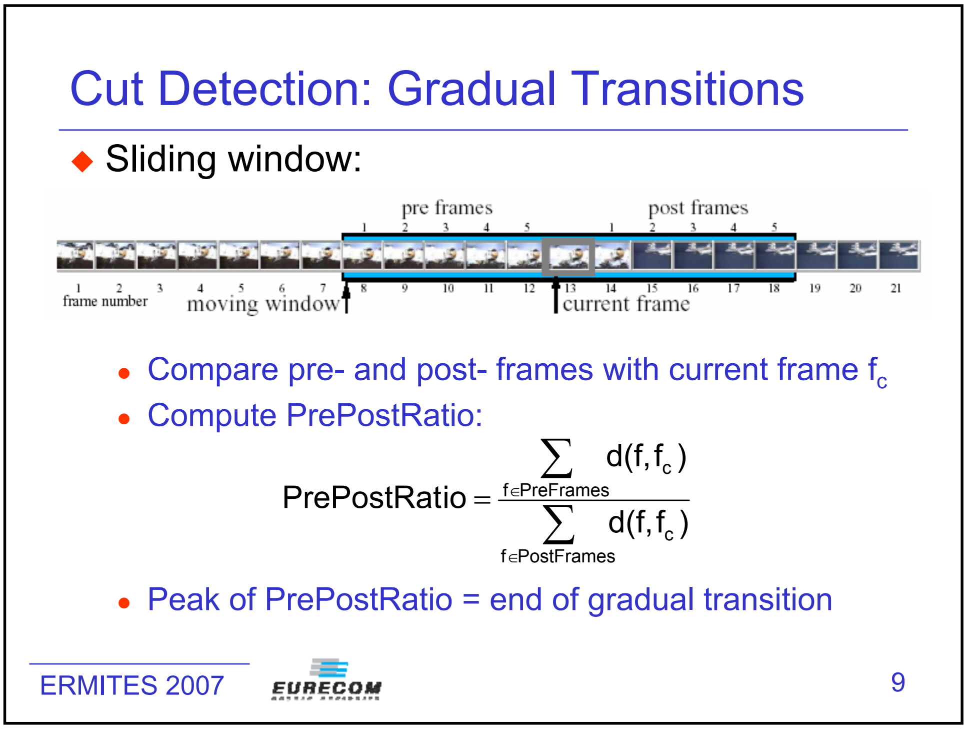
- Basic idea:
  - Measure distance $d(\mathbf{I}_t,\mathbf{I}_{t+1})$ between consecutive frames
  - Detect cut if distance is greater than threshold:

    $d(\mathbf{I}_t,\mathbf{I}_{t+1}) \geq \theta$
- Common distance: color histogram

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



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



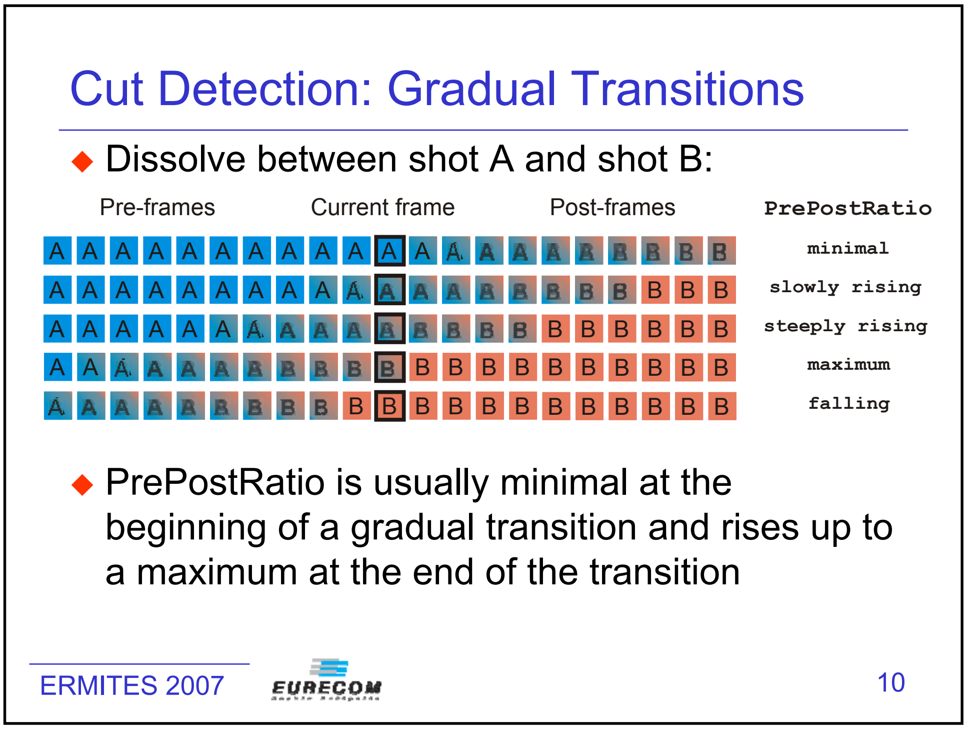
Cut Detection: Gradual Transitions

- Sliding window:

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

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

  - Peak of PrePostRatio = end of gradual transition



Cut Detection: Gradual Transitions

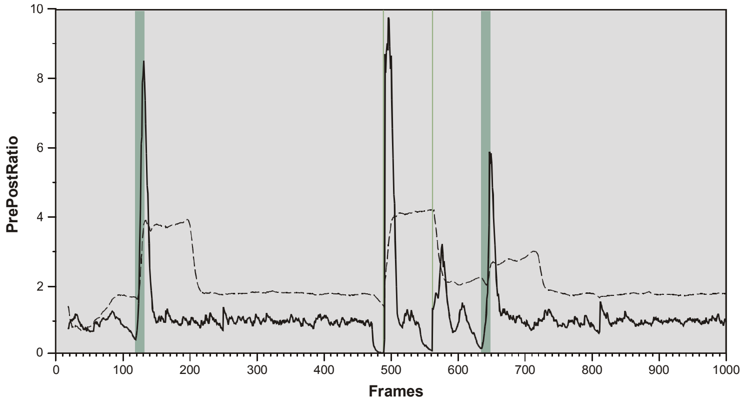
- Dissolve between shot A and shot B:

			Pre-frames							Current frame							Post-frames				PrePostRatio
A	A	A	A	A	A	A	A	A	A	A	A	Á	A	A	A	B	B	B	B	B	minimal
A	A	A	A	A	A	A	A	A	Á	A	A	A	A	B	B	B	B	B	B	B	slowly rising
A	A	A	A	A	A	Á	A	A	A	B	B	B	B	B	B	B	B	B	B	B	steeply rising
A	A	Á	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	maximum
Á	A	A	A	A	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	falling

- 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

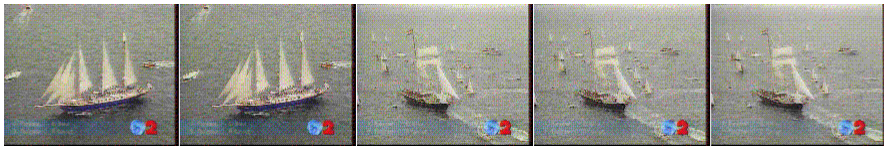
- Example of PrePostRatio curve



- two short gradual transitions and two cuts

Cut detection: difficult cases

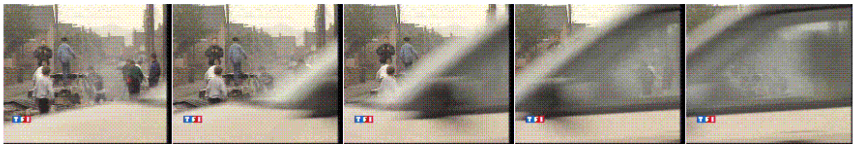
- Similar environment



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

Cut detection: difficult cases

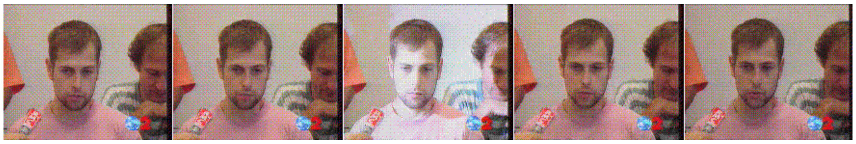
- Fast movement of large object



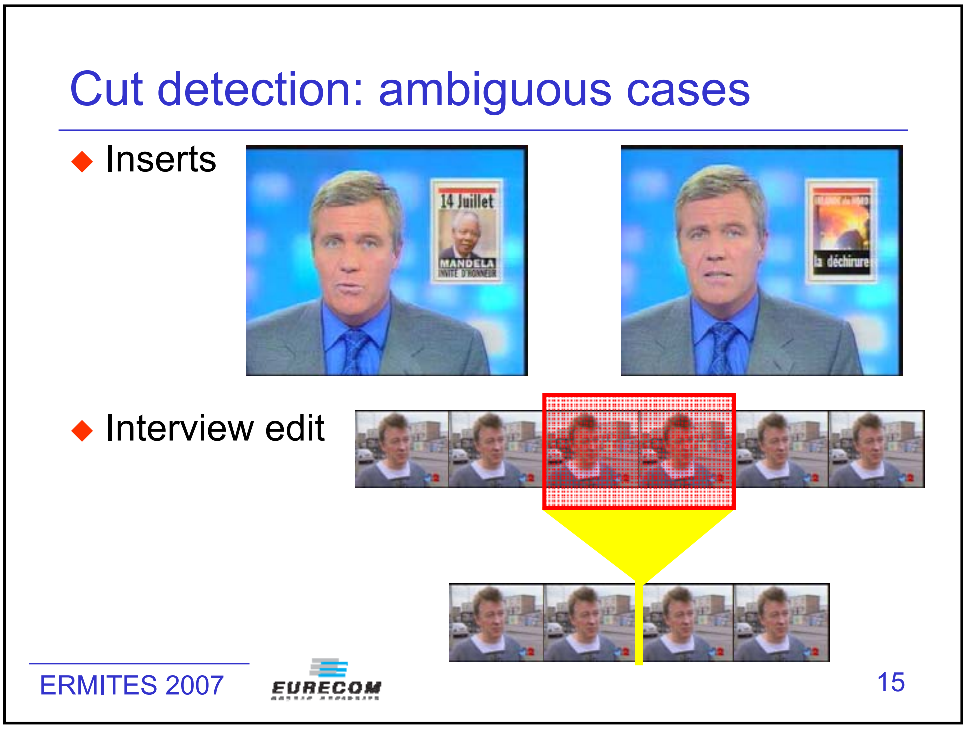
- Can be confused with wipe
- Shot can be over-segmented

Cut detection: difficult cases

- Sudden change in illumination



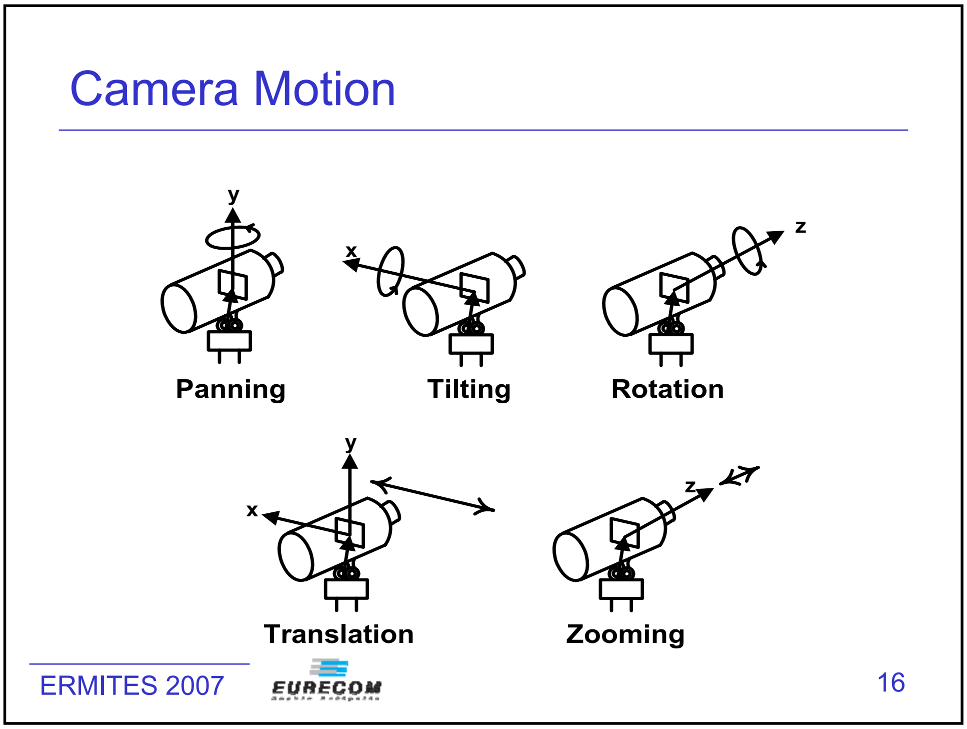
- Sudden modification of colors
- Also the case in explosions, etc…
- Shot can be over-segmented



Cut detection: ambiguous cases

- Inserts
- Interview edit

ERMITES 2007 EURECOM 15

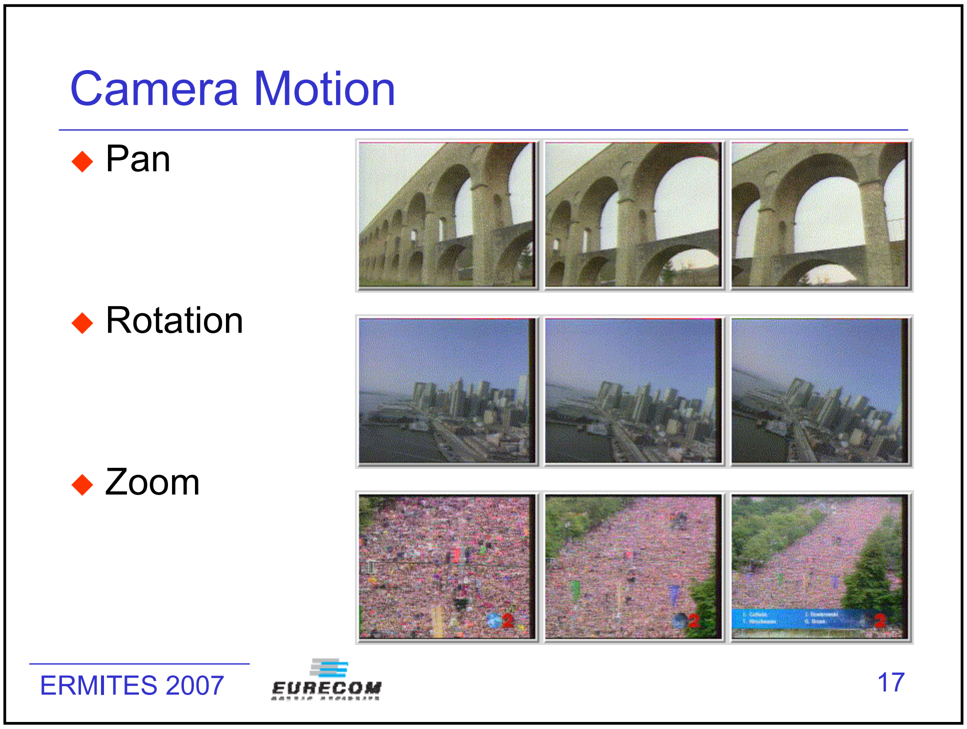


Camera Motion

Panning Tilting Rotation

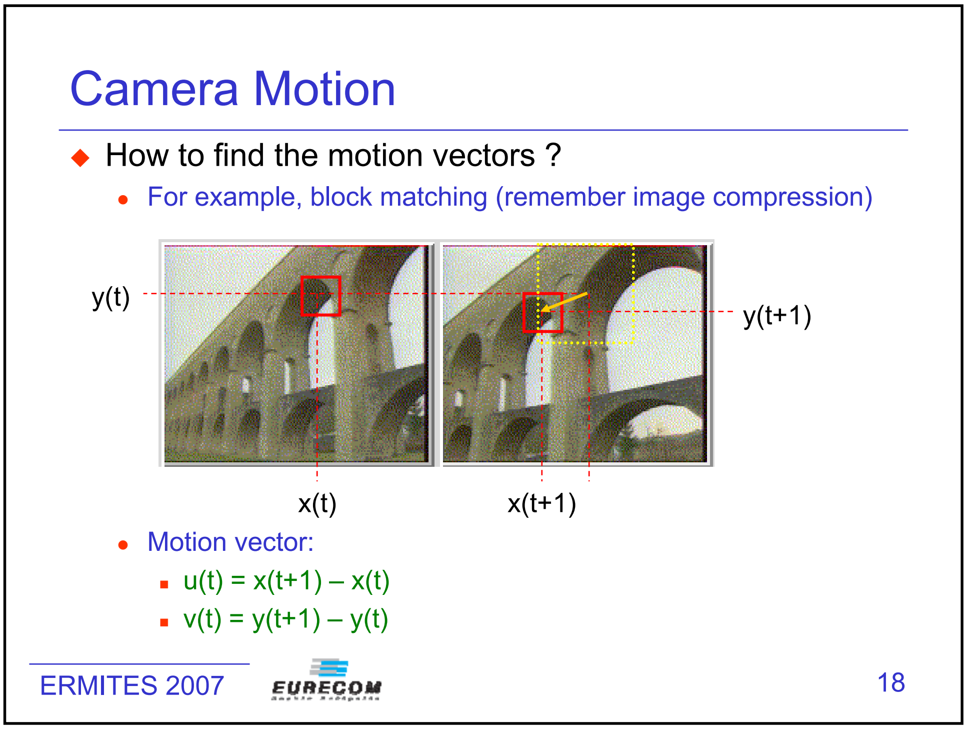
Translation Zooming

ERMITES 2007 EURECOM 16



Camera Motion

- Pan
- Rotation
- Zoom



Camera Motion

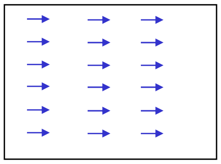
- How to find the motion vectors ?
  - For example, block matching (remember image compression)

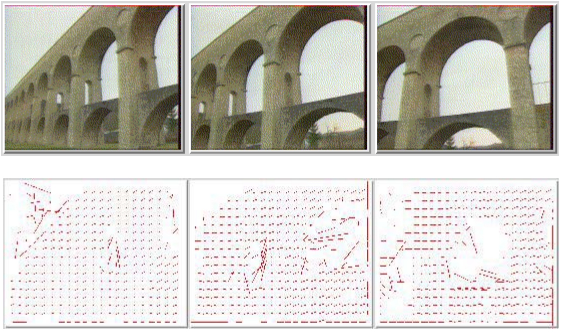
y(t)
y(t+1)
x(t)
x(t+1)

  - Motion vector:
    - u(t) = x(t+1) – x(t)
    - v(t) = y(t+1) – y(t)

Camera Motion : Pan

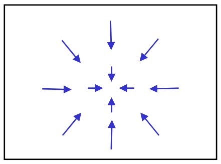
- Ideal Real

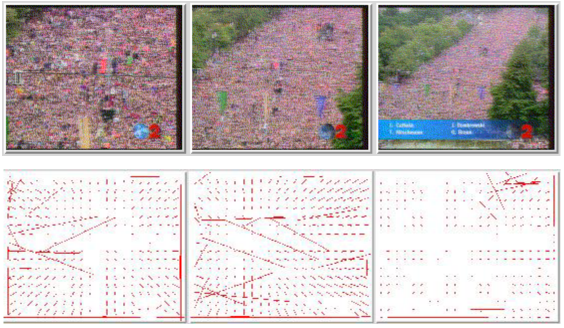


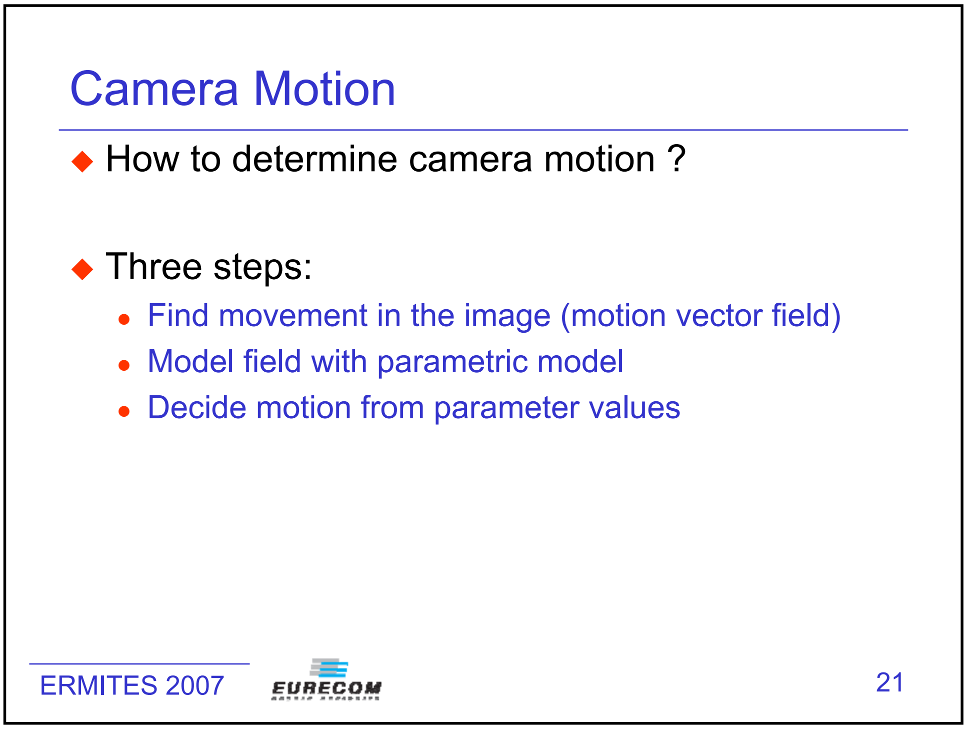


Camera Motion : Zoom In

- Ideal Real

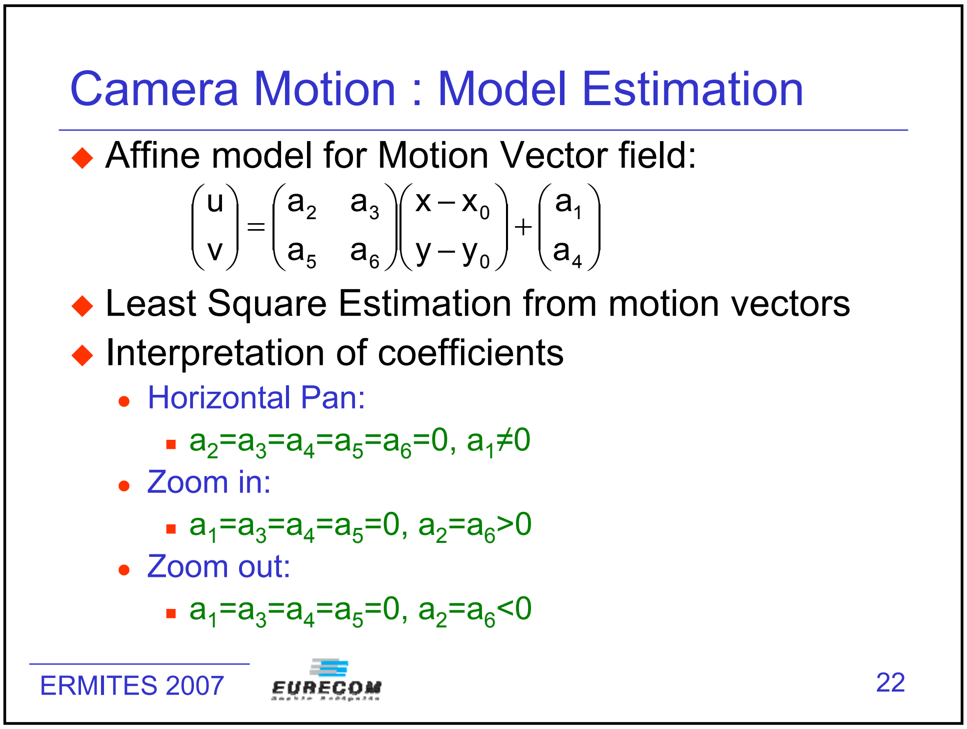






Camera Motion

- How to determine camera motion ?
- Three steps:
  - Find movement in the image (motion vector field)
  - Model field with parametric model
  - Decide motion from parameter values



Camera Motion : Model Estimation

- Affine model for Motion Vector field:

$$\begin{pmatrix}u\\v\end{pmatrix}=\begin{pmatrix}a_2 & a_3\\a_5 & a_6\end{pmatrix}\begin{pmatrix}x-x_0\\y-y_0\end{pmatrix}+\begin{pmatrix}a_1\\a_4\end{pmatrix}$$

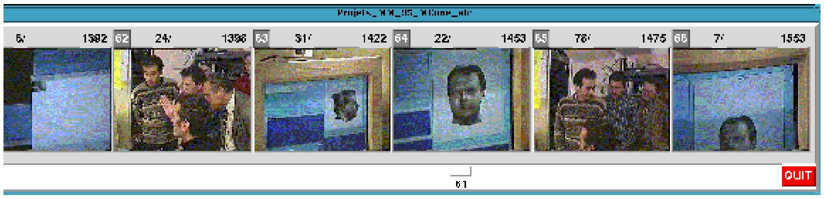
- Least Square Estimation from motion vectors
- Interpretation of coefficients
  - Horizontal Pan:
    - $a_2=a_3=a_4=a_5=a_6=0$, $a_1\neq0$
  - Zoom in:
    - $a_1=a_3=a_4=a_5=0$, $a_2=a_6>0$
  - Zoom out:
    - $a_1=a_3=a_4=a_5=0$, $a_2=a_6<0$

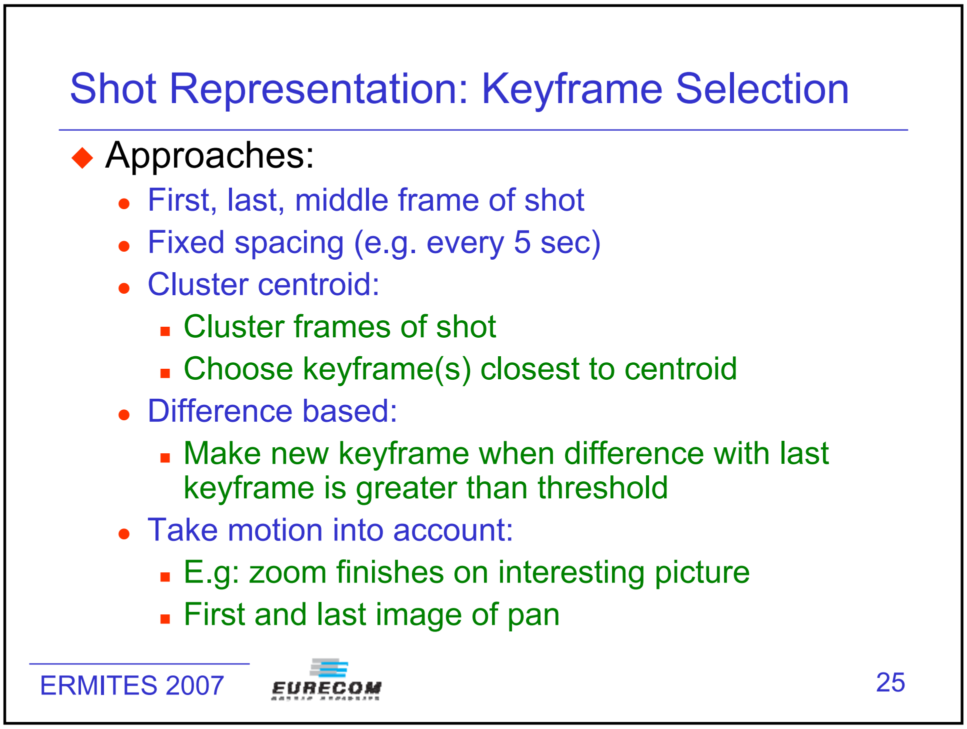
Camera Motion : Other Models

Model	Coordinate transformation	Parameters	Degree of freedoms
Translation	$\mathbf{x}' = \mathbf{x} + \mathbf{b}$	$\mathbf{b} \in \mathbb{R}^2$	2
Rigid	$\mathbf{x}' = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \mathbf{x} + \mathbf{b}$	$\theta \in [0, 2\pi), \mathbf{b} \in \mathbb{R}^2$	3
Affine	$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$	$\mathbf{A} \in \mathbb{R}^{2\times 2}, \mathbf{b} \in \mathbb{R}^2$	6
Bilinear	$x' = q_0xy + q_1x + q_2y + q_3$ $y' = q_4xy + q_5x + q_6y + q_7$	$q_i \in \mathbb{R}$ where $i = 0, 1, 2, \ldots, 7$	8
Projective	$x' = \frac{\mathbf{Ax+b}}{\mathbf{c}^T\mathbf{x}+1}$	$\mathbf{A} \in \mathbb{R}^{2\times 2}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^2$	8
Pseudo-perspective	$x' = q_0x + q_1y + q_2 + q_3x^2 + q_4xy$ $y' = q_5x + q_6y + q_7 + q_8xy + q_9y^2$	$q_i \in \mathbb{R}$ where $i = 0, 1, 2, \ldots, 9$	10

Shot Representation: Keyframes

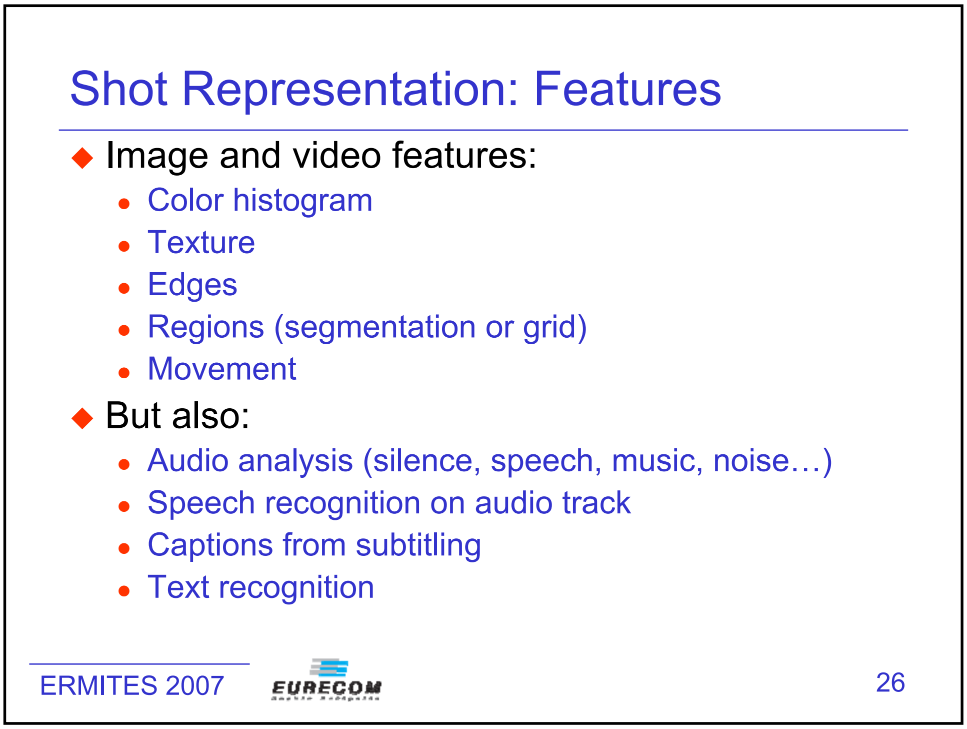
- It is interesting to represent a shot with a single image: keyframe





Shot Representation: Keyframe Selection

- Approaches:
  - First, last, middle frame of shot
  - Fixed spacing (e.g. every 5 sec)
  - Cluster centroid:
    - Cluster frames of shot
    - Choose keyframe(s) closest to centroid
  - Difference based:
    - Make new keyframe when difference with last keyframe is greater than threshold
  - Take motion into account:
    - E.g: zoom finishes on interesting picture
    - First and last image of pan

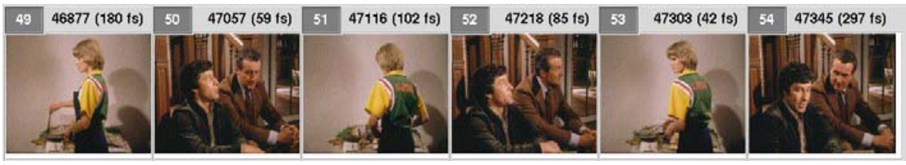


Shot Representation: Features

- Image and video features:
  - Color histogram
  - Texture
  - Edges
  - Regions (segmentation or grid)
  - Movement
- But also:
  - Audio analysis (silence, speech, music, noise…)
  - Speech recognition on audio track
  - Captions from subtitling
  - Text recognition

Scene Segmentation

- A scene is a sequence of shots with similar topic, action and/or location
  - Alternance of similar shots



  - Same location, similar content

Scene Segmentation

- Idea:
  - Similar shots close in time belong to the same scene
- Algorithm:
  - Clustering with temporal distance:

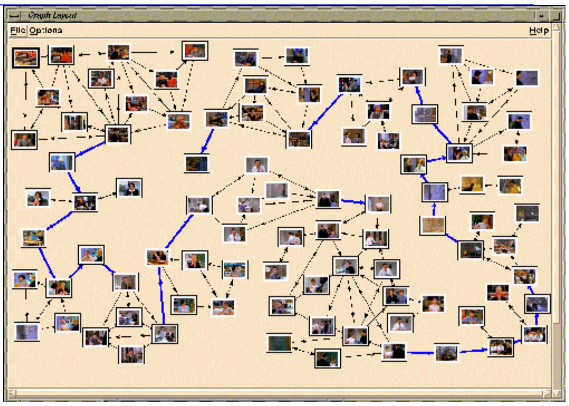
$$d(i,j) = \begin{cases} 100 - s(i,j) \times W(i,j) & \text{if } |i-j| < T \\ \text{infinity} & \text{else} \end{cases}$$

with:

s(i,j)	similarity of shots i and j (normalized to 100)
W(i,j)	temporal weight function (vanishes for |i-j|=T)

Scene Transition Graphs

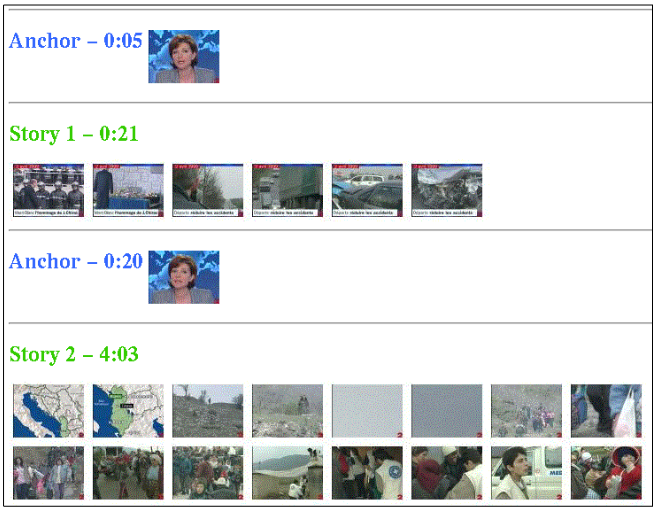
Yeung & Yeo, Princeton



- Cluster shots and build graph of shot sequences
- Scenes are islands connected by a single arc in graph

TV News parsing

- Detect anchor
- split into stories



Anchor – 0:05

Story 1 – 0:21

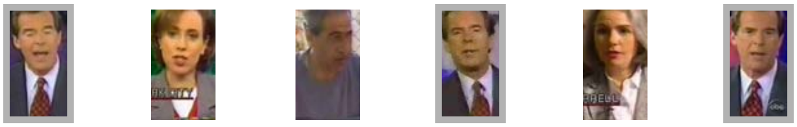
Anchor – 0:20

Story 2 – 4:03

Anchor Detection (UCF)

- Problem:
  - Same anchor person on different backgrounds

- Face Detection
  - Based on color
- Person Detection
  - Extend face region and cluster



TrecVid

EURECOM Sophia Antipolis

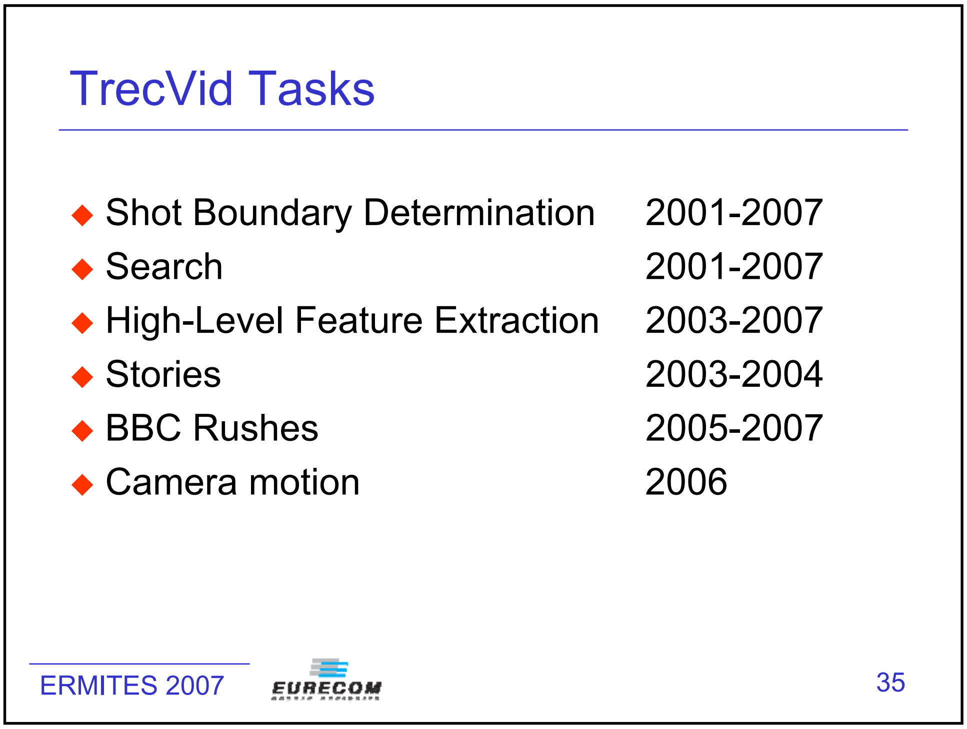
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

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 2006

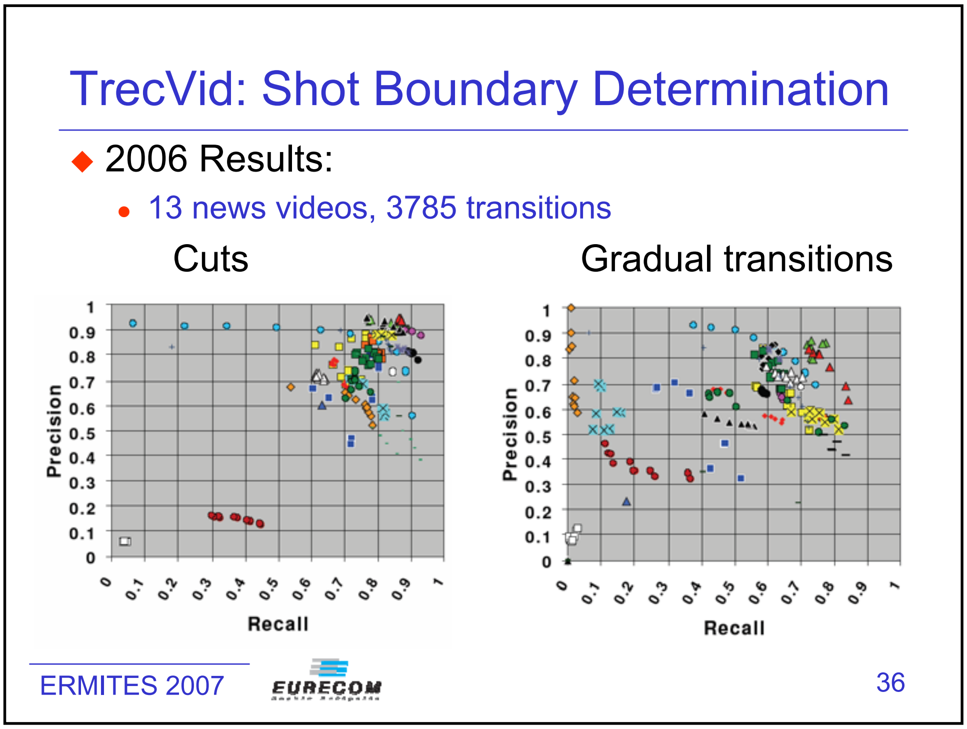
TrecVid Data

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)
	50/50	BBC Rushes



TrecVid Tasks

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



TrecVid: Shot Boundary Determination

- 2006 Results:
  - 13 news videos, 3785 transitions

Cuts

Gradual transitions

TrecVid: High Level Feature Extraction

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

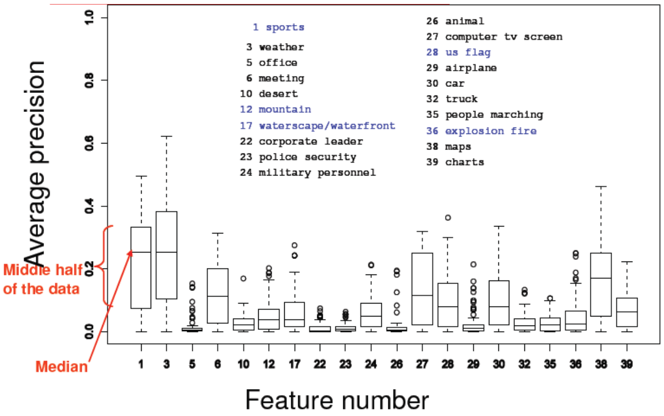
Sports
Entertainment
Weather
Court
Office
Meeting
Studio
Outdoor
Building
Desert
Vegetation
Mountain
Road

Sky
Snow
Urban
Waterscape_Waterfront
Crowd
Face
Person
Government-Leader
Corporate-Leader
Police_Security
Military
Prisoner
Animal

Computer_TV-screen
Flag-US
Airplane
Car
Bus
Truck
Boat_Ship
Walking_Running
People-Marching
Explosion_Fire
Natural-Disaster
Maps
Charts

TrecVid: High Level Feature Extraction

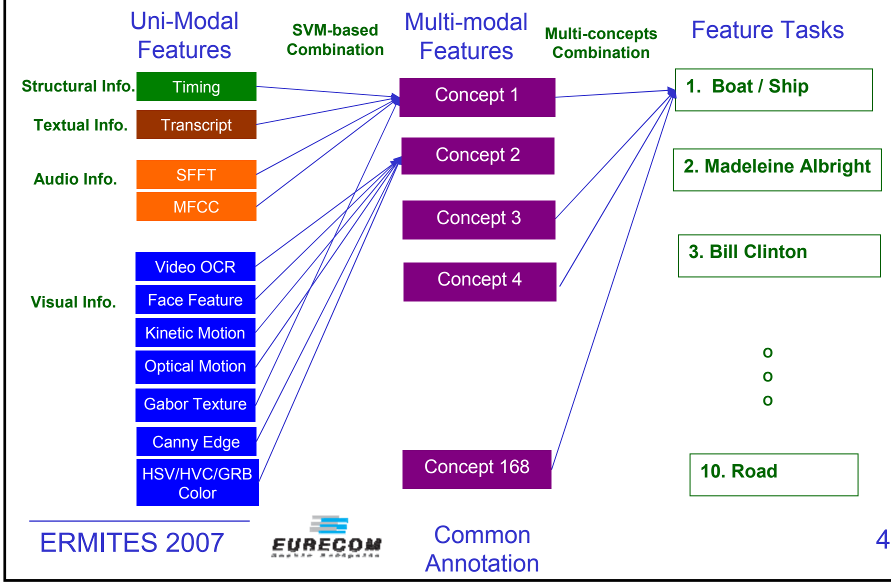
- 2006: 20 concepts evaluated
- Overall results:



TrecVid: High Level Feature Extraction

- Example: CMU system (2005):
  - Low level features extraction
    - Image
    - Audio
    - Motion
    - Detectors (face & text)
  - Elementary Concept classifiers (168 concepts)
    - SVM (Support Vector Machines)
  - Multi-Classifier fusion

TrecVid : CMU Architecture



Uni-Modal Features

SVM-based Combination

Multi-modal Features

Multi-concepts Combination

Feature Tasks

Structural Info. Timing

Textual Info. Transcript

Audio Info.

SFFT

MFCC

Visual Info.

Video OCR

Face Feature

Kinetic Motion

Optical Motion

Gabor Texture

Canny Edge

HSV/HVC/GRB Color

Concept 1

Concept 2

Concept 3

Concept 4

Concept 168

1. Boat / Ship
2. Madeleine Albright
3. Bill Clinton

o
o
o

10. Road

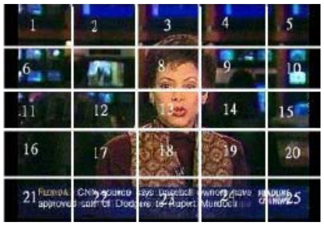
Common Annotation

TrecVid : CMU Low level features

- 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 Image features

- 5 by 5 grids for key-frame per shot
- Color histogram
  - 5 by 5, 125 bins color histogram
  - HSV, HVC, and RGB color space
  - 3125 dimensions (5*5*125)
  - row-wise grids
- Texture
  - Six orientated Gabor filters
- Edge
  - Canny edge detector, 8 orientations



TrecVid : CMU Audio & Motion

- Every 20 msecs (512 windows at 44100 HZ sampling rate)
  - FFT – Short Time Fourier Transform
  - MFCC – Mel-Frequency cepstral coefficients
  - SFFT – simplified FFT
- Kinetic energy
  - Capture the pixel difference between frames
- Mpeg motion
  - Mpeg motion vector extracted from p-frame
- Optical flow
  - Capture optical flow in each grid

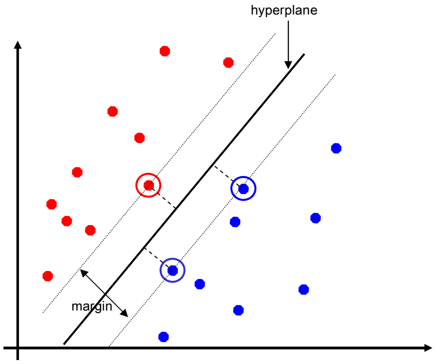
TrecVid : CMU Detector features

- Face detector
  - Detecting faces in the images

- VOCR detector
  - Detecting and recognizing VOCR

TrecVid : CMU SVM Classifier

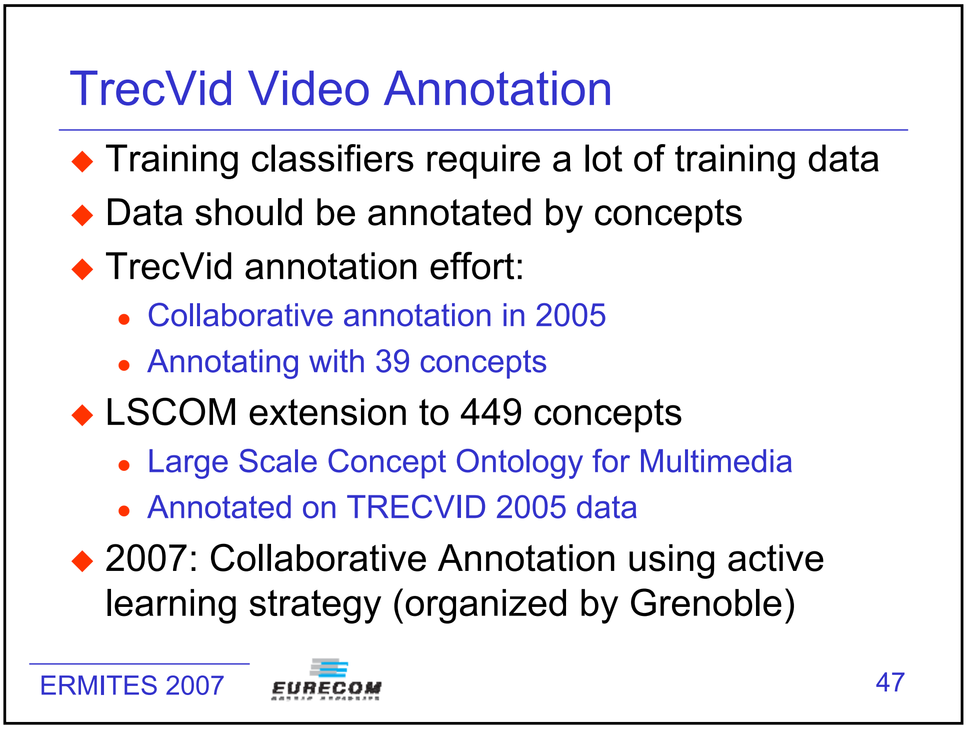
- Binary classifier
- Constructed from training data
- Linear separator with highest margin in space with Kernel distance



TrecVid : CMU Multi-concepts Combination

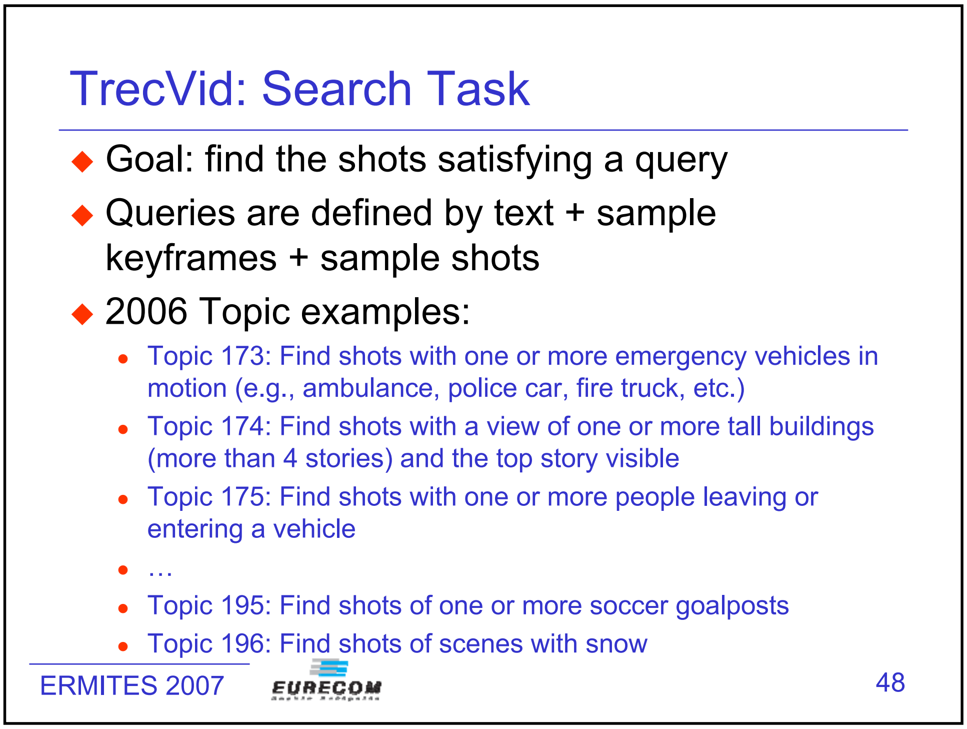
- Bayesian Networks from 168 common annotation concepts
- Combine 4 most related concepts with target concept

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 Video Annotation

- Training classifiers require a lot of training data
- Data should be annotated by concepts
- TrecVid annotation effort:
  - Collaborative annotation in 2005
  - Annotating with 39 concepts
- LSCOM extension to 449 concepts
  - Large Scale Concept Ontology for Multimedia
  - Annotated on TRECVID 2005 data
- 2007: Collaborative Annotation using active learning strategy (organized by Grenoble)

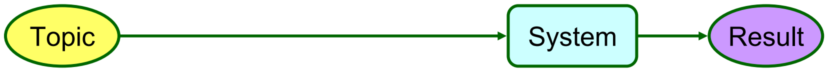


TrecVid: Search Task

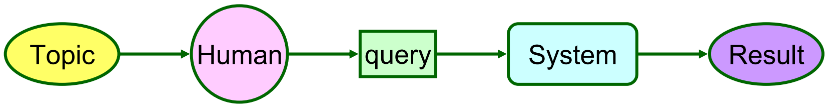
- Goal: find the shots satisfying a query
- Queries are defined by text + sample keyframes + sample shots
- 2006 Topic examples:
  - Topic 173: Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)
  - Topic 174: Find shots with a view of one or more 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

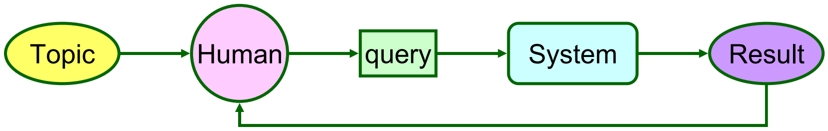
- 3 types of experiments:
  - Automatic:



  - Human-assisted:

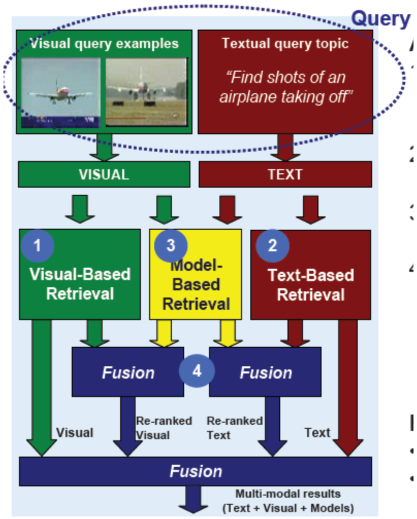


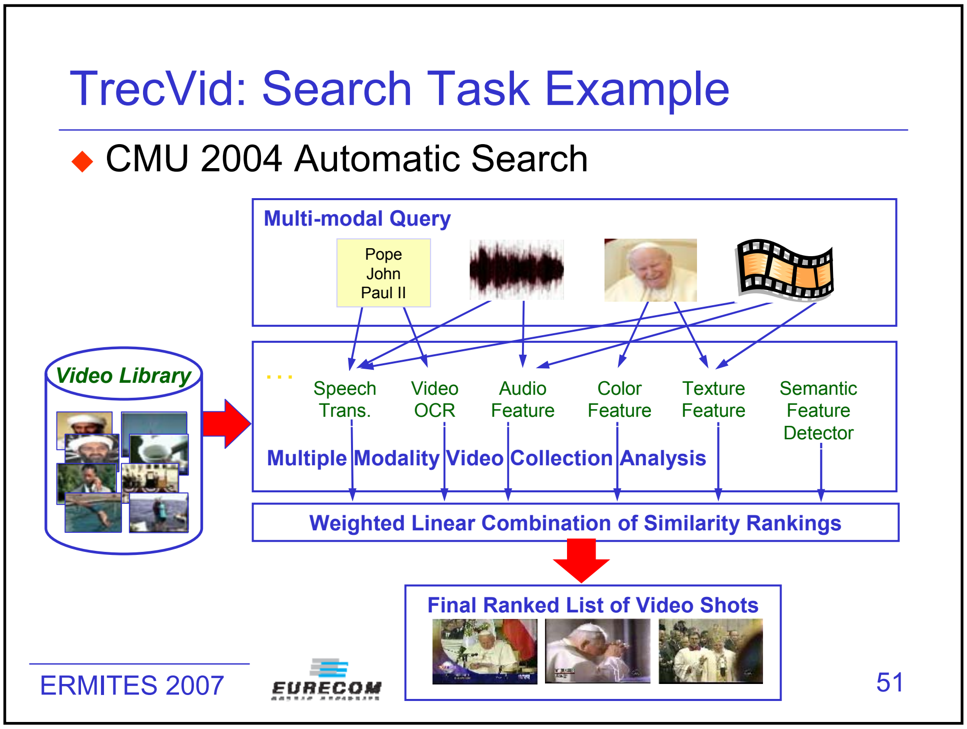
  - Interactive:



TrecVid: Search Task Example

- IBM 2005 Automatic Search
- 1.Visual-based:
  - light-weight learning (discriminative and nearest neighbor modeling)
- 2.Text-based:
  - automatic query expansion
- 3.Model-based:
  - automatic query-to-model mapping & weighting
- 4.Fusion:
  - Query-independent
  - Statistical normalization (visual)
  - Rank normalization (text)
  - Model-based re-ranking (text & visual)





TrecVid: Search Task Example

- CMU 2004 Automatic Search

Multi-modal Query

Pope John Paul II

Video Library

. . .

Speech Trans.

Video OCR

Audio Feature

Color Feature

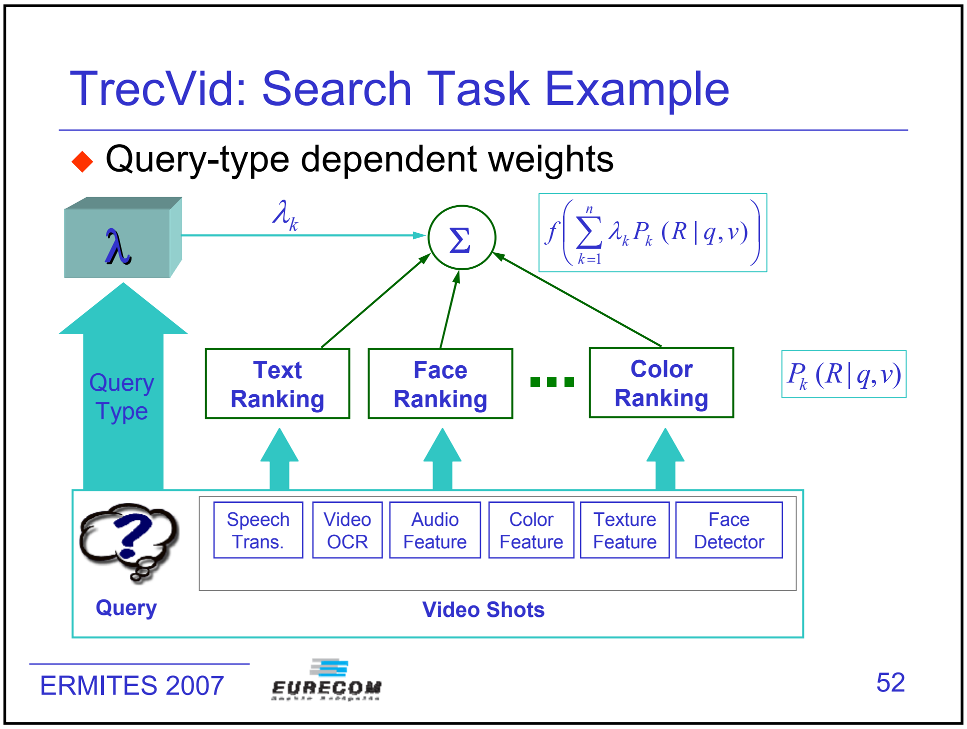
Texture Feature

Semantic Feature Detector

Multiple Modality Video Collection Analysis

Weighted Linear Combination of Similarity Rankings

Final Ranked List of Video Shots



TrecVid: Search Task Example

- Query-type dependent weights

λ

λ_k

Σ

$f\left(\sum_{k=1}^{n}\lambda_k P_k(R|q,v)\right)$

Query Type

Text Ranking

Face Ranking

...

Color Ranking

$P_k(R|q,v)$

Speech Trans.

Video OCR

Audio Feature

Color Feature

Texture Feature

Face Detector

Query

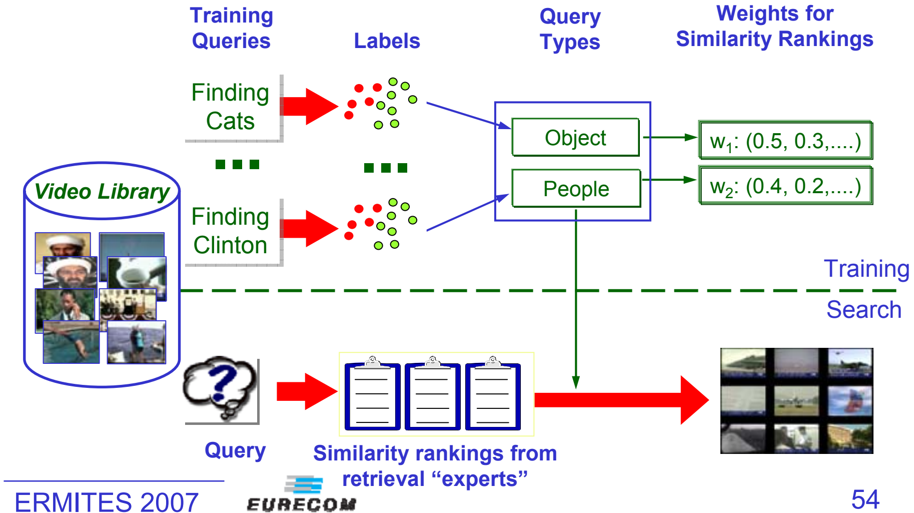
Video Shots

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"

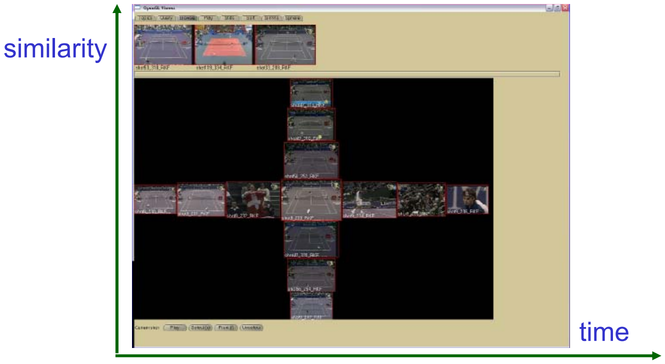
TrecVid: Search Task Example

- Learning Query-type dependent weights



TrecVid: Search Task Example

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



TrecVid: BBC Rushes

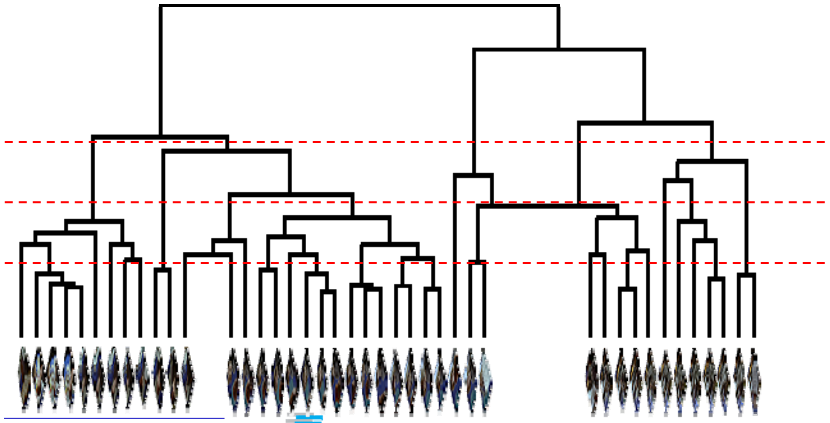
- 2005: no task
- 2006: organize, no evaluation
- 2007: summarize, evaluation
  - List of topics and events as ground truth
  - 10 topics picked randomly
  - Evaluator watches summary and counts topics present

TrecVid: BBC Rushes

- Topics for video MRS044493:
  - man from distance riding bicycle approach camera
  - camera pans man with knapsack riding bicycle
  - House
  - zooming in
  - bald man looking out of window
  - bald man at window
  - closeup at boy with fair hair looking out of window,only head visible
  - zooming in boy with fair hair at window
  - closeup at boy with fair hair at window looking at camera
  - door
  - bald man walking out of the house and a woman followed behind quarelling
  - woman pull out bald's man shirt entering storeroom
  - bald man and woman talking in store room
  - bald man went out the storeroom
  - bald man talking with man in dark overalls
  - bald man enter store room
  - bald man open cupboard and pull out things
  - bald man shake woman's cheek and talking
  - bald man went out store room and woman followed
  - 3 people standing talking one facing camera
  - …

Eurecom Summarization

- Hierarchical clustering of 1 sec video segments
  - Level to be adjusted based on video content



Eurecom Summarization

- Detection of event redundancy:
  - If two shots contain replays of the same event, they should contain (almost) the same sequences of clusters

Cluster 1	Cluster 1	Cluster 2	Cluster 3	Cluster 3

Eurecom Summarization

- Shot selection principle:
  - Find a set of shots which best covers the set of clusters
- Shot selection algorithm:
  - Greedy selection:
    - Make all clusters initially active
    - Find shot which contains most active clusters
    - Select this shot and make its clusters inactive
    - Iterate

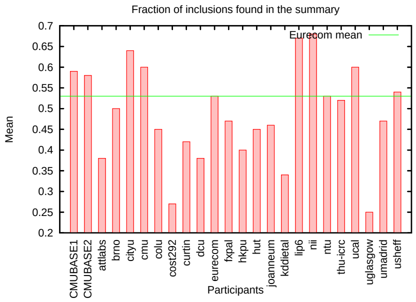
Eurecom Summarization

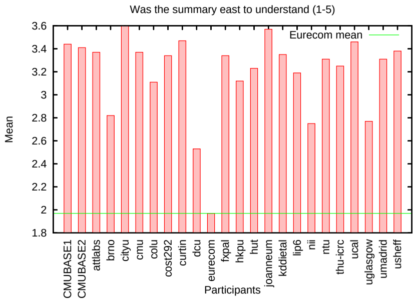
- Additional refinements:
  - Cluster value:
    - Based on activity and presence of faces
  - Dynamic acceleration:
    - Accelerate a shot to minimize static content

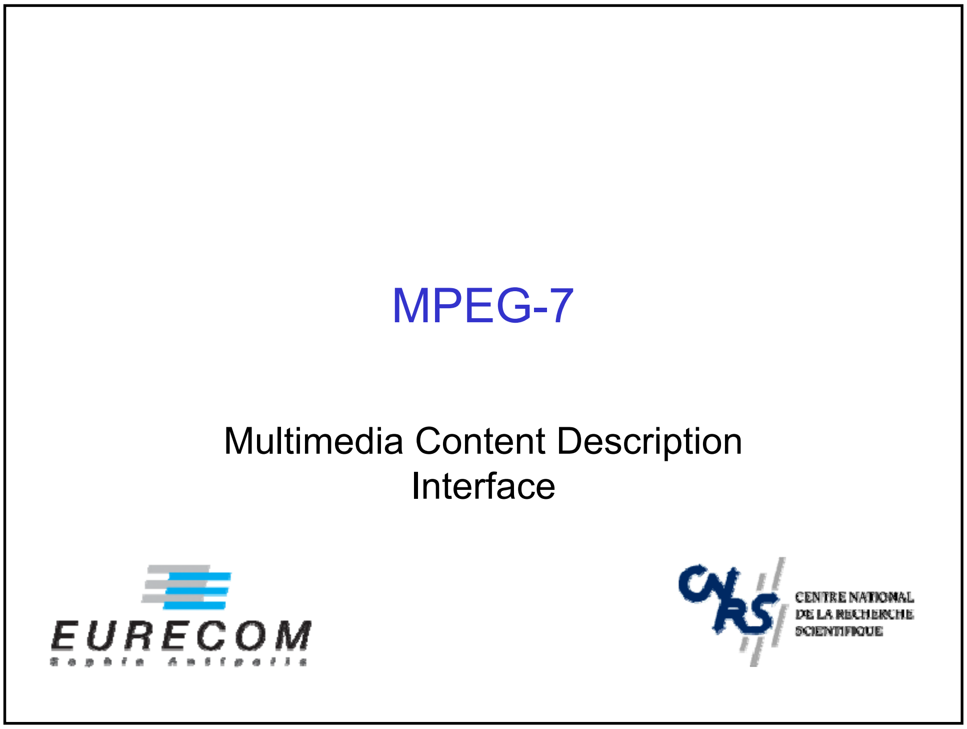
  - Split-scren display
    - Maximize information displayed by time unit
    - Group shots by 4

Eurecom Summarization

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

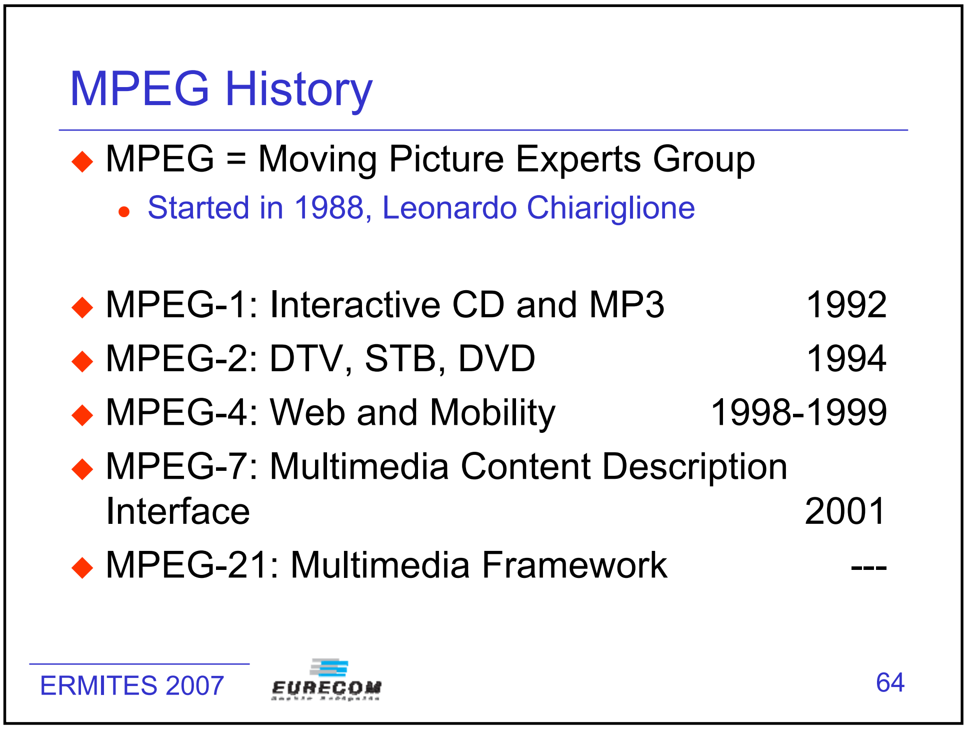






MPEG-7

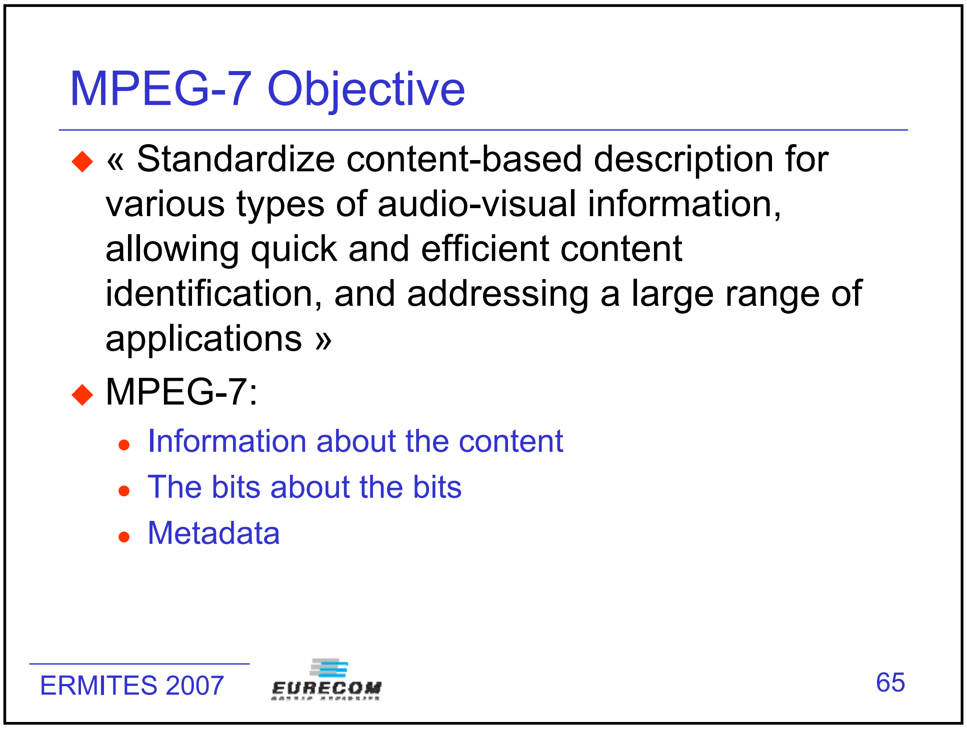
Multimedia Content Description
Interface



MPEG History

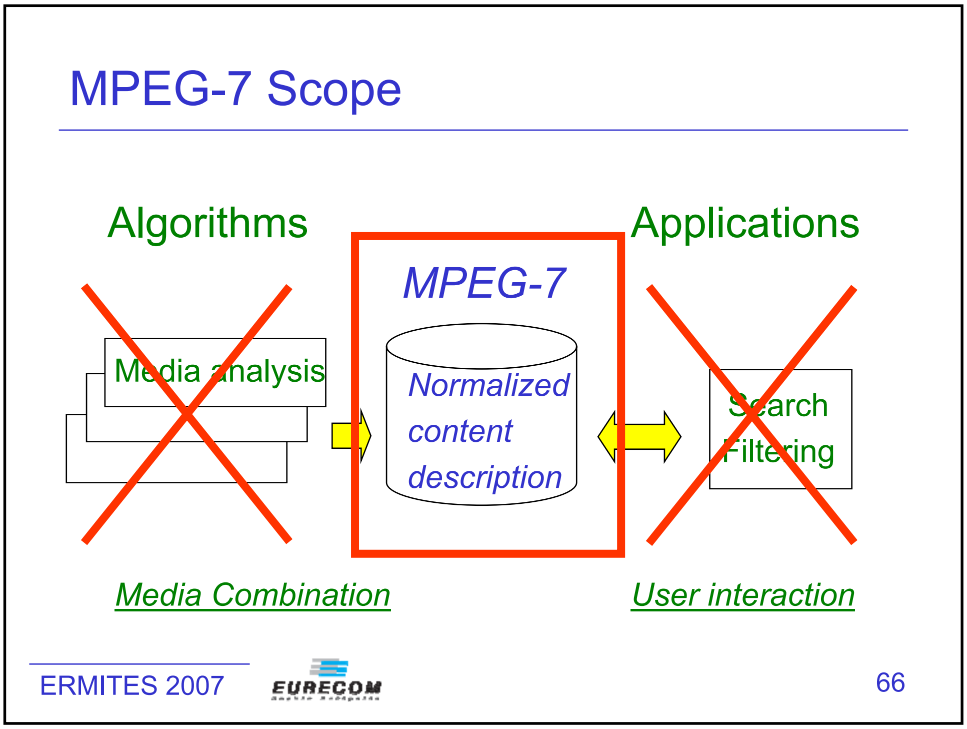
- MPEG = Moving Picture Experts Group
  - Started in 1988, Leonardo Chiariglione

- 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

Algorithms

Applications

MPEG-7

Media analysis

Normalized content description

Search Filtering

Media Combination

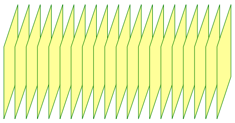
User interaction

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

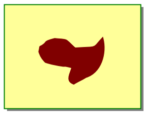
Low level Audio Visual descriptors

Video segments



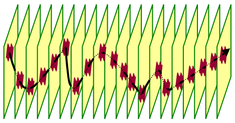
- Color
- Camera motion
- Motion activity
- Mosaic

Still regions



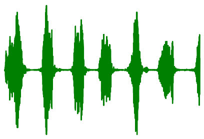
- Color
- Shape
- Position
- Texture

Moving regions



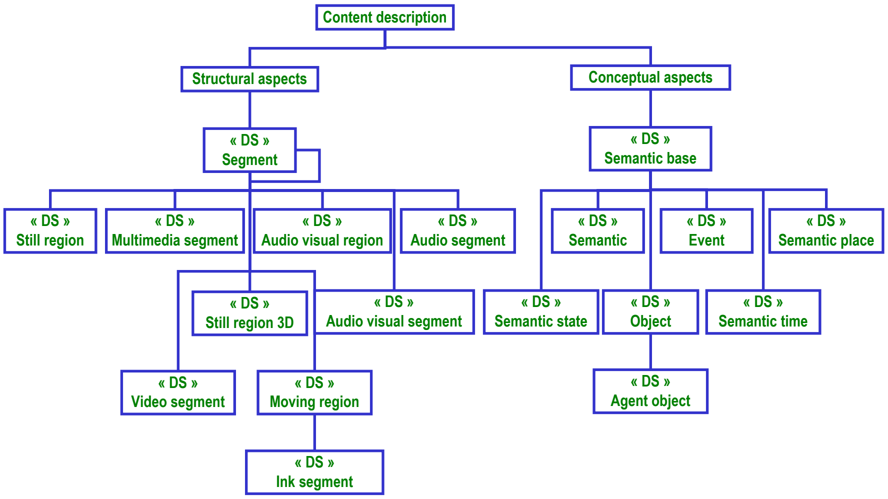
- Color
- Motion trajectory
- Parametric motion
- Spatio-temporal shape

Audio segments



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

Multimedia DS: Content Description



Content description

Structural aspects

Conceptual aspects

« DS »
Segment

« DS »
Semantic base

« DS »
Still region

« DS »
Multimedia segment

« DS »
Audio visual region

« DS »
Audio segment

« DS »
Semantic

« DS »
Event

« DS »
Semantic place

« DS »
Still region 3D

« DS »
Audio visual segment

« DS »
Semantic state

« DS »
Object

« DS »
Semantic time

« DS »
Video segment

« DS »
Moving region

« DS »
Agent object

« DS »
Ink segment

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