On the Synergy between Indexing and Compression Representations for Video Sequences

Ph.D. Dissertation

Javier Ruiz Hidalgo

Dept. Teoria Senyal i Comunicacions
Universitat Politècnica de Catalunya

14 de Junio de 2007

Departament de Teoria
del Senyal i Comunicacions

UNIVERSITAT POLITÈCNICA DE CATALUNYA

sy·ner·gy, noun

The working together of two or more things to produce an effect greater than the sum of their individual effects.

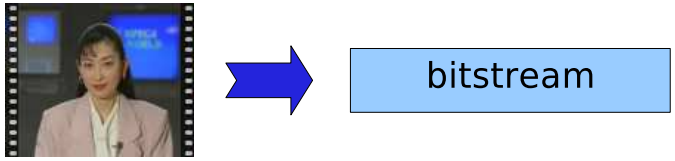
Presentation Outline

Motivation

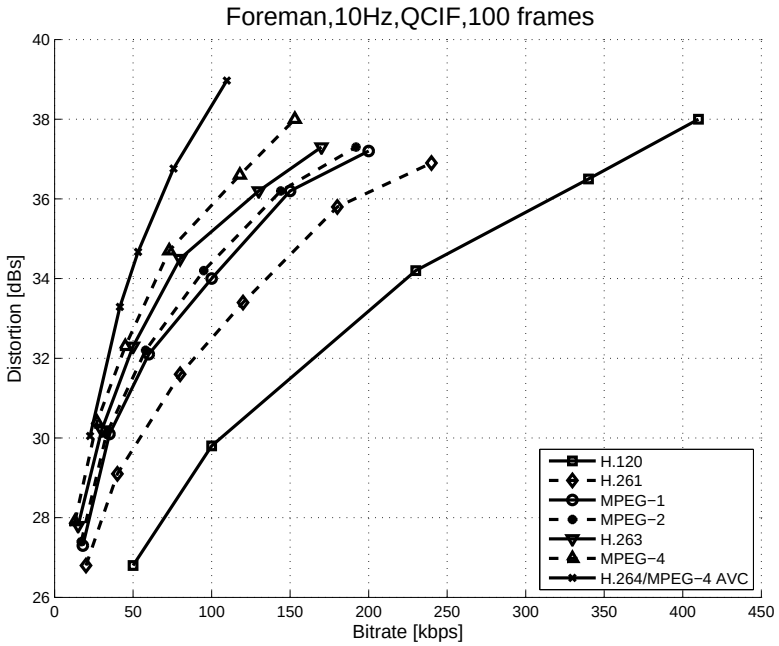
- Large amount of digital audiovisual content
- Provide Functionalities:
  - **Storage**
  - **Transmission in limited bandwidth networks**
  - **Browsing**
  - **Summarizing**
  - **Search & Retrieval**
  - **Consumption**
- **Compression** and **Indexing** Representations

Compression Representation Definition

Compression Representations aim at depicting the digital audiovisual content in a compact manner using the minimum number of bits possible and allowing its visualization



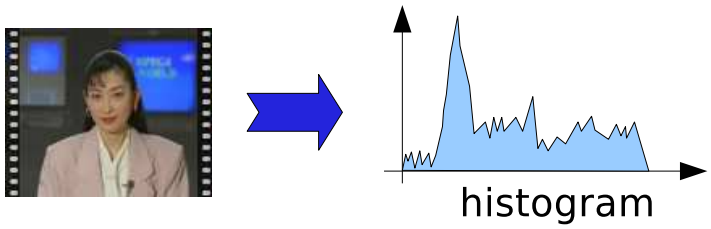
Standard Compression Representations



- **1988** **H.120**: Frame difference coding
- **1991** **H.261**: Motion compensation
- **1993** **MPEG-1**: B-frames, 1/2 pel
- **1994** **MPEG-2**: Prediction modes, VLC tables
- **1996** **H.263**: Variable block sizes, arithmetic entropy encoder
- **2000** **MPEG-4**: Direct mode, 1/4 pel, mosaic coding
- **2003** **H.264**: Long term prediction, integer transform, CABAC

Indexing Representation Definition

Indexing Representations, also known as **metadata**, refer to all data structures, descriptors or extra information that are extracted from the content in order to provide functionalities of summarizing, browsing, search & retrieval, etc.

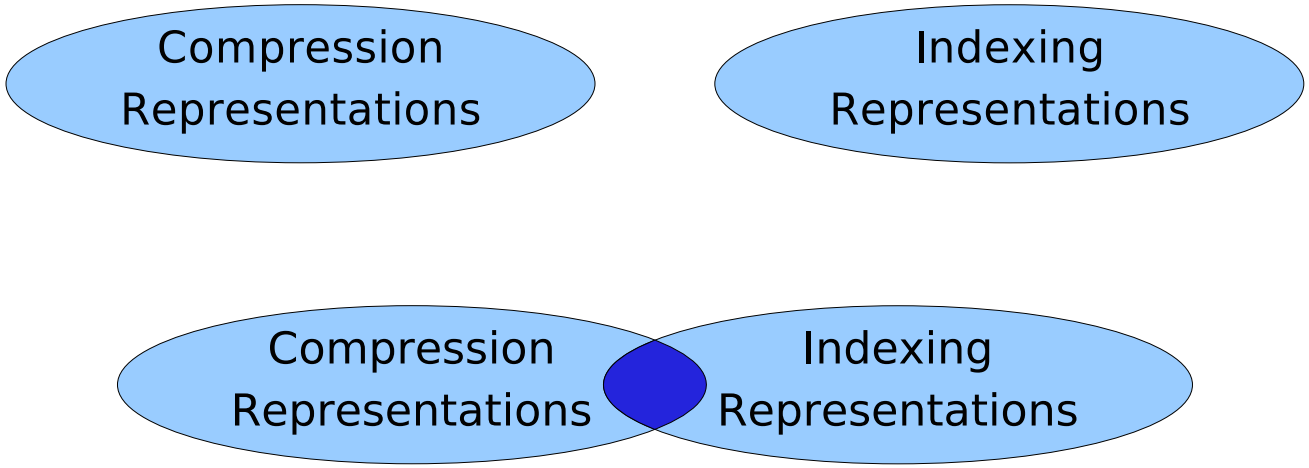


Standard Indexing Representations

- **2001** **SMPTE** Metadata Dictionary: Definition of registered set of pairs (tags,values) to describe multimedia content
- **2003** **MPEG-7** Multimedia content description interface: Provides descriptions to allow the searching, indexing, browsing or accessing to multimedia content
- **2003+** **MPEG-21**: Normative framework for multimedia delivery and consumption

Objectives

- Compression and indexing representations considered as **separate** representations providing different functionalities
- **Can these representations be combined?**
  - Ideally: One representation able to provide all functionalities efficiently
  - Actually: Using one representation to improve the other or sharing data between both representations



Compression Representations Useful for Indexing

- Most of audiovisual content is present in compressed form
- Exploit compression representations
  - Extract indexing representations from compressed ones
  - Share data between representations, example of motion vectors:
    - Very useful in compression representations
    - Also useful for indexing sequences by motion type
- In this work, use mosaics (sprites):
  - Extracted for compression representations (sprite coding scheme MPEG-4)
  - Used to represent background information and to extract foreground regions of video shots

Key-frame Shot Representation

- Several key-frames are used to represent the shot

Content Based Shot Representation

- Using key-frames to represent shots is not always efficient
  - Shots with very high internal motion
  - Shots of very long duration
- Abstraction layers that represent the content of the video scene, separating background and foreground information

Mosaics in Shot Representations

- Mosaics represent the background information
- **How to represent the foreground objects of the shot?**

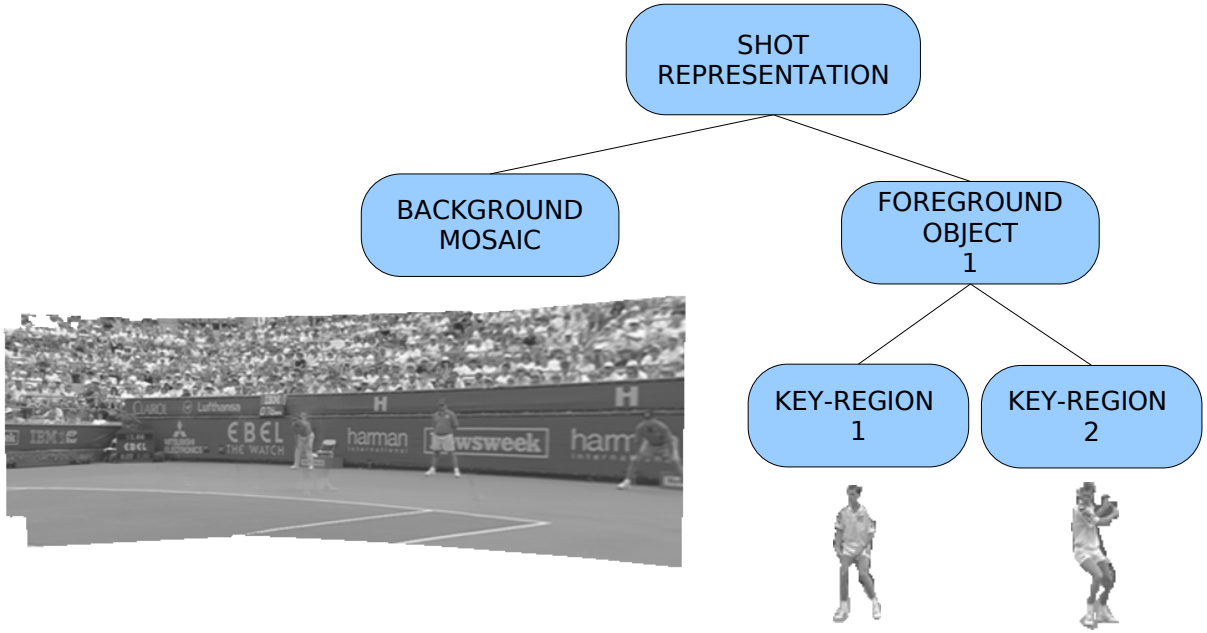
1996 **Shawney**: Use several mosaics

1998 **Irani**: Synopsis mosaics

2003 **Fusiello**: Snapshots of foreground regions

Proposed Shot Representation

- Mosaic represents the background
- Foreground objects are represented by a set of key-regions
  - Deformations of non-rigid foreground objects are characterized by one or more key-regions



Key-regions

Based on three images:

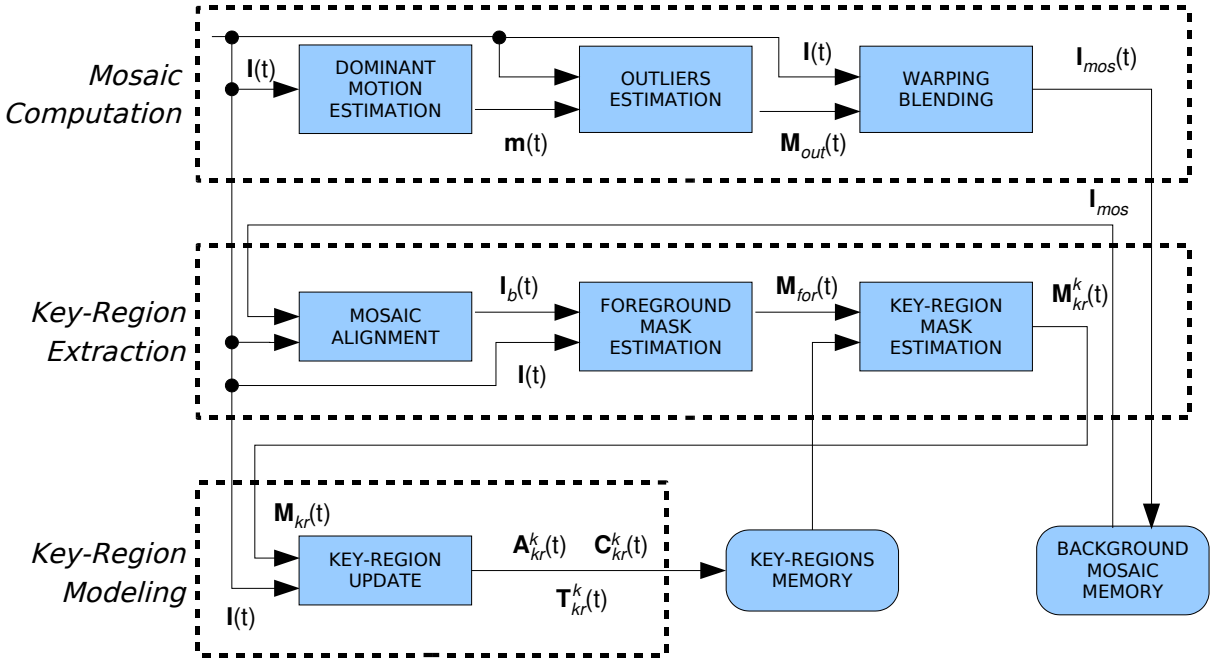
- Appearance image ($\mathbf{A}_{kr}$): Number of times a pixel has been estimated as belonging to the key-region
- Contours image ($\mathbf{C}_{kr}$): Confidence of the key-region contour over time
- Texture image ($\mathbf{T}_{kr}$): Overall texture of the key-region

$\mathbf{A}_{kr}$ $\mathbf{C}_{kr}$ $\mathbf{T}_{kr}$

General Scheme

The shot representation algorithm works in a two pass approach:

1. Mosaic is created
2. Key-regions are extracted and grouped into foreground objects



Mosaic Alignment

Final mosaic (created using all frames in the shot) and motion information for each frame (inverse warping parameters) used to create a background representation of the scene at each time instant t

- ✓ Foreground regions are removed
- × Two pass approach. Inappropriate for real-time applications

Foreground Mask Estimation

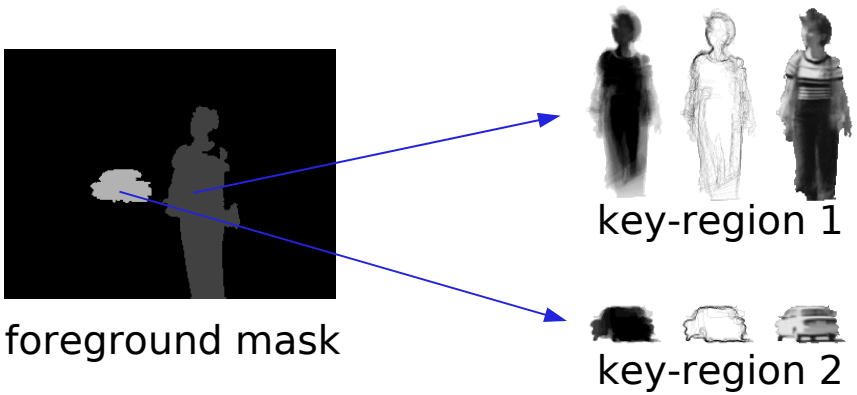
Watershed algorithm applied on the difference between original and background information

1. Calculate difference: $\mathbf{D}(t) = |\mathbf{I}(t) - \mathbf{I}_b(t)|$
2. Watershed markers for foreground and background regions: $\epsilon_s \{\mathbf{D}(t) < t_{back}\} \cup \epsilon_s \{\mathbf{D}(t) > t_{for}\}$
3. Watershed gradient: $\mathcal{G}\{\mathbf{I}(t) - \mathbf{I}_b(t)\} \cdot (\mathcal{G}\{\mathbf{I}(t) - \mathbf{I}(t-1)\} \vee t_0)$
4. Clean result and group connected regions

Key-Region Mask Estimation (I)

Foreground masks are improved based on previous estimation of key-regions

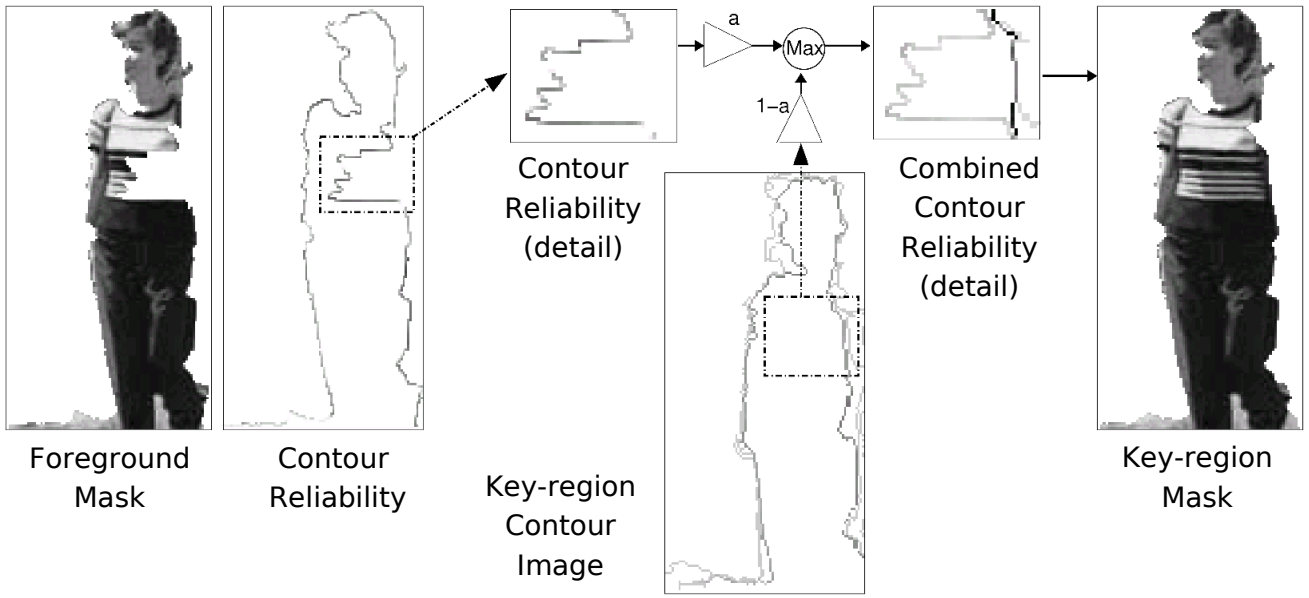
- Previous key-region association based on overlap



- Motion compensation between foreground and key-regions

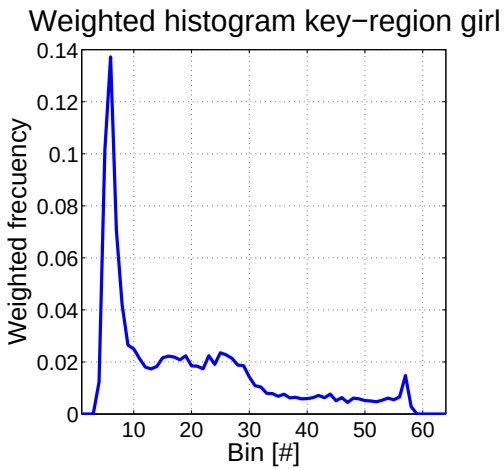
Key-Region Mask Estimation (II)

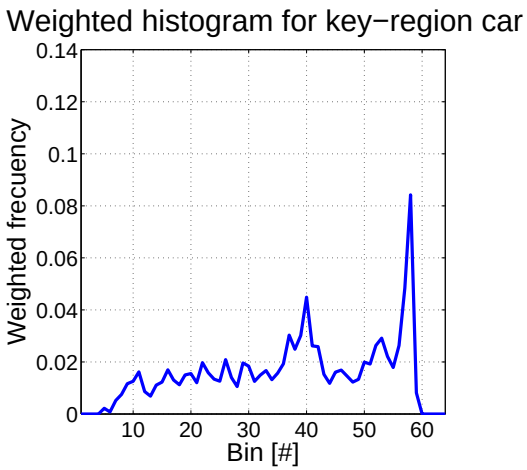
- Improve foreground masks:
  1. Contour reliability of compensated foreground mask $\hat{\mathbf{M}}_{for}(t)$:
     $\mathbf{C}_r(t) = \mathcal{C}\left\{\mathcal{G}\{\hat{\mathbf{I}}(t)\}, \hat{\mathbf{M}}_{for}(t)\right\}$
  2. Combine contour with key-region contour image $\mathbf{C}_{kr}(t-1)$:
     $a \cdot \mathbf{C}_r(t) \vee (1-a) \cdot \mathbf{C}_{kr}(t-1)$
  3. Key-region mask estimation using watershed



Key-Region Modelling and Grouping

- Appearance, contours and texture images of the key-region are updated with extracted information on a frame by frame basis
- After all key-regions are extracted, they are grouped to form foreground objects
  1. Weighted histograms extracted for each key-region texture image
  2. Intersections between weighted histograms used to group foreground objects

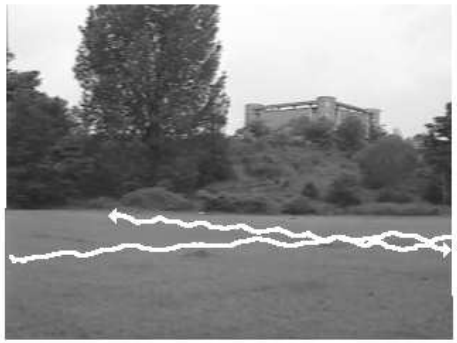




Shot Representation Results (I)

Composed of synopsis mosaic and foreground objects

ETRI_od_B sequence



Mosaic and trajectories

Foreground objects

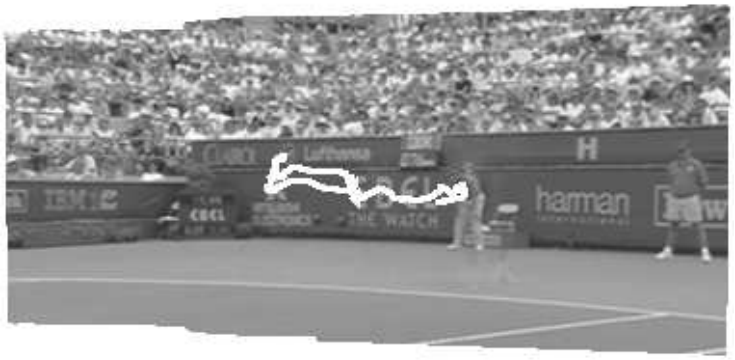
NHKvideo7 sequence

Mosaic and trajectories

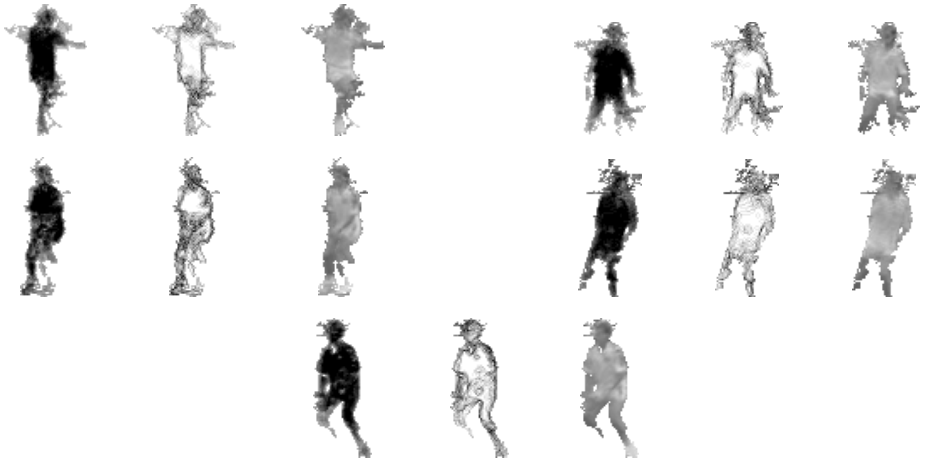
Foreground objects

Shot Representation Results (II)

Stefan sequence



Mosaic and trajectories



Foreground object 1

Foreground object 2

Conclusions

- ✓ New indexing representation to summarize video shots
- ✓ Exploit compression representation by sharing mosaics
- ✓ Morphological operators to improve background information in mosaics
- ✓ Create key-regions that represent foreground objects
- × Inappropriate for real-time applications

Motivation

- Large amount of digital audiovisual content
- In the near future, digital content will have to be indexed to be useful
- Indexing representations will be available for compression representations
- **Are indexing representations also useful to improve compression representations?**

Compression Representations Using Indexing Representations

Few contributions reported in the literature:

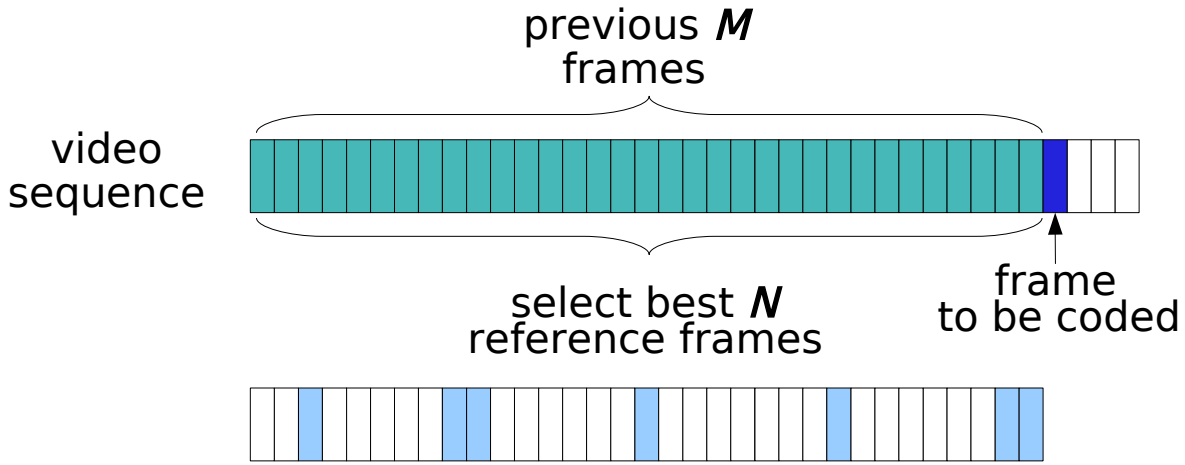
- **2002** **Smolic**: Use of the MPEG-7 Parametric Motion descriptor to improve the motion estimation/compensation step of H.264
- **2003** **Ndjiki-Nya**: MPEG-7 texture descriptors to identify and separately encode textured blocks in H.264
- **2003** **Torres**: Encoding of faces using the MPEG-7 Face Recognition descriptor

Proposed techniques

- Video transition coding
  - Encode video transitions exploiting information from MPEG-7 Video Editing Segment
  - Bitrate reductions up to 90% for short transitions and 65% for longer transitions
- Frame type selection
  - Selection of the optimum *GoP* structure based on MPEG-7 Motion Activity descriptor
  - Similar bitrates for the best fixed GoP structure possible for each sequence
- Long term reference frame selection
- Video segment shuffling

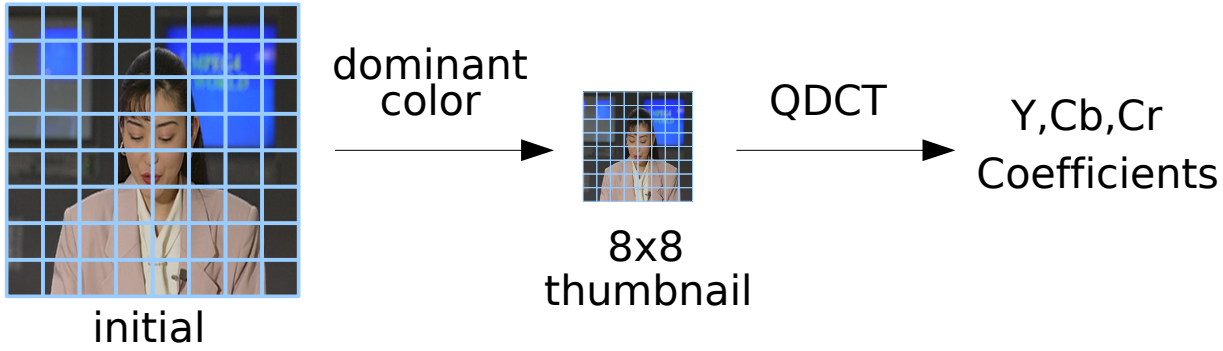
Motivation

- Exploit temporal redundancy: Long term temporal prediction (use N reference frames)
  - N limited due to computational complexity and bitrate increase
- Reformulate motion estimation in a **search & retrieval** approach:
  - Indexing representations may be employed
  - Number of possible reference frames can be greatly increased



Selected Indexing Representation

- MPEG-7 Color Layout descriptor
  - Describe the spatial distribution of colors in an image
  - Simple, easy to compute with similarity measures
- Extraction algorithm, 18 coefficients selected
- Similarity between color layouts based on a weighted Euclidean distance between QDCT coefficients



Coding Scheme

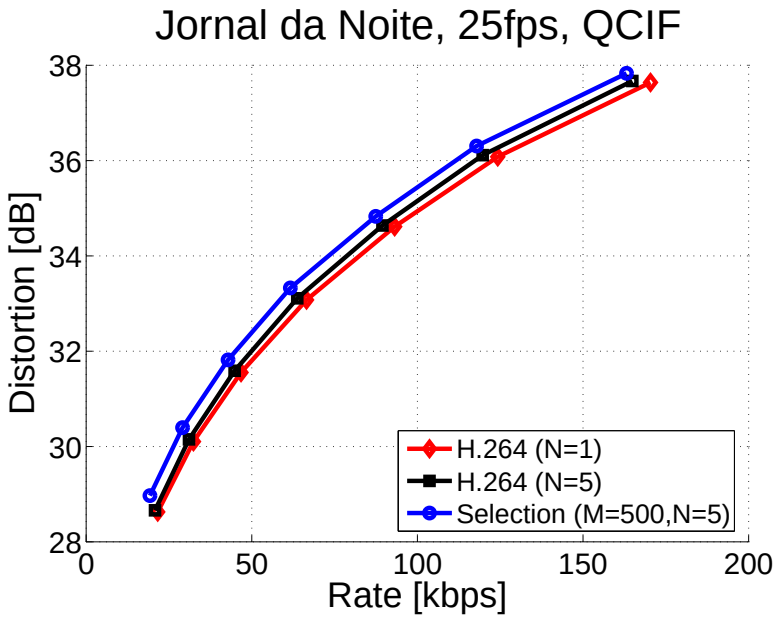
The process to fill the H.264 long term prediction buffer with N reference frames:

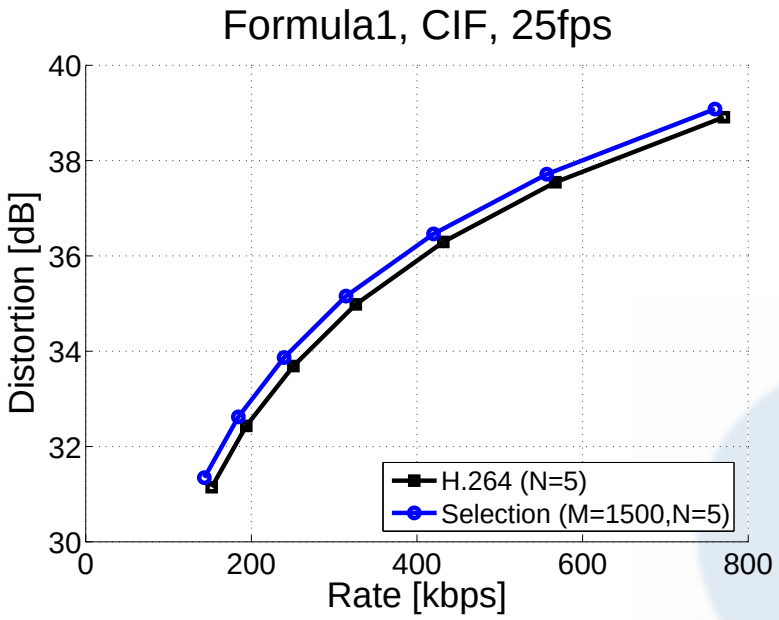
1. Obtain and store Color Layout description for current frame
2. Compare current color layout with M previous Color Layout descriptors
3. Fill the long term prediction buffer with the N coded frames with greater similarity
4. Perform the motion estimation step on the modified long term prediction buffer

Experimental Results (I)

Indexing representation is available at the decoder

- Implemented in H.264
- Greater gains with sequences corresponding to news, interviews, sports, etc.





Experimental Results (II)

Indexing representation **not** available at the decoder

- Sending the color layouts amounts to 15 bits per frame (compressed)
- The same information is sent at all bitrates and resolutions

Test	Data H.264 [kbps]	Savings [kbps]	Indexing Representation [kbps]
Jornal da noite (QCIF)	20.82	1.55 (7.4%)	0.350 (1.68%)
	164.90	1.74 (1.1%)	0.350 (0.21%)
Formula1 (QCIF)	61.03	3.30 (5.4%)	0.425 (0.70%)
	345.93	5.80 (1.7%)	0.425 (0.12%)

Conclusions

Long Term Reference Frame Selection using MPEG-7 Color Layout descriptors:

- ✓ Greater gains with sequences with repeating shots or with objects re-appearing in the scene, up to 15% bitrate savings at the same visual quality
- ✓ Less than 1% increase in computational complexity
- × Increase in memory complexity of the decoder

Motivation

- Exploit temporal redundancy in video sequences:
  - The visual order of frames in a video may not be the most efficient for coding
  - *B* frames prove that altering the visual order is more efficient
  - Generally, the order is changed only a small amount of frames
- For non-real time applications, interesting to re-arrange the entire sequence globally
- Video segment shuffling proposes to use a high level indexing representation to create a new coding order for the entire sequence

Selected Indexing Representation (I)

The MPEG-7 Video Segment description scheme:

- Provides mechanisms to produce a *Table of Contents* for accessing or an *Index* for searching video content
- Includes information such as: Location, media, creation, semantic, structural relations
- Segments can be described with other MPEG-7 descriptors

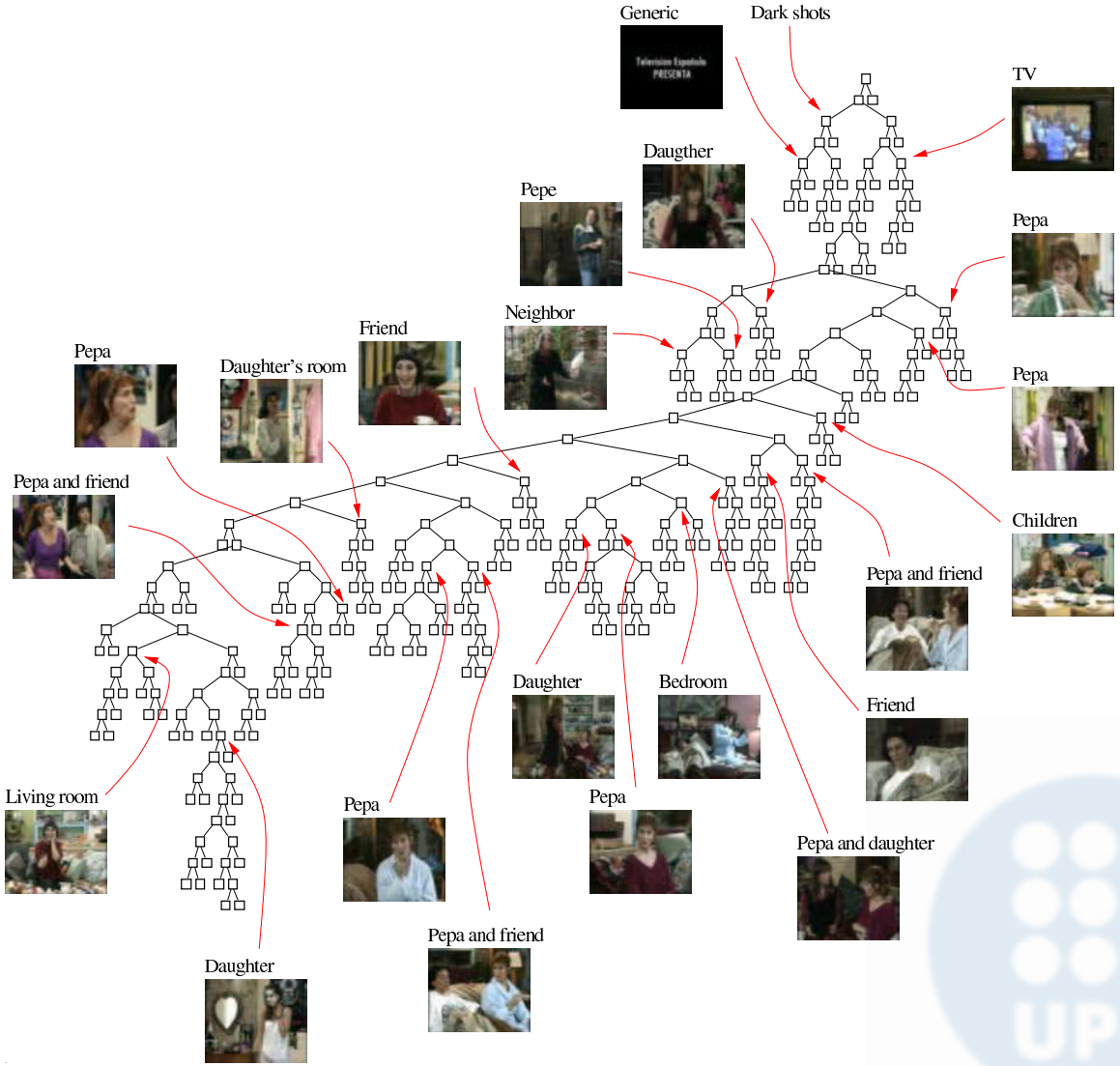
Selected Indexing Representation (II)

An indexing Representation based on MPEG-7 Video Segment can be the **Binary Partition Tree** (BPT)

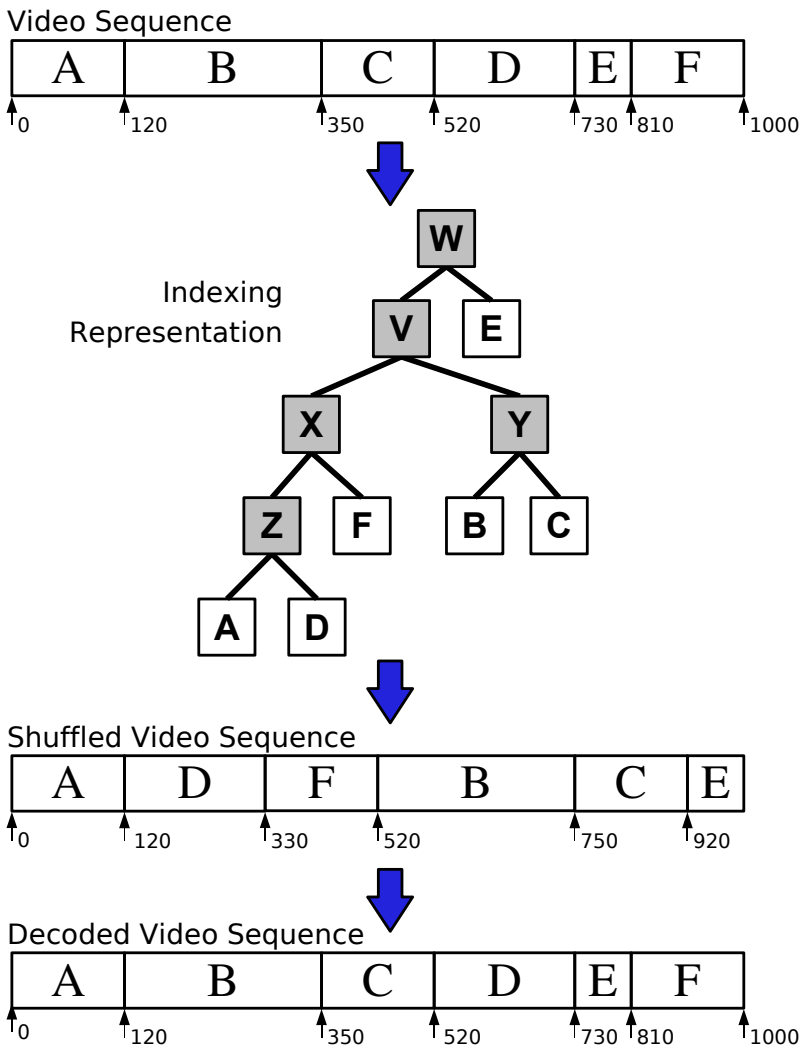
- Developed at the UPC image group
- Hierarchical representation of images or video sequences from an initial partition
- Initial partition: single frames, groups of L frames or complete shots
- Similarity criterion is employed to merge video segments
- Only two video segments are merged at a time

Example Video Segment representation

- *Drama* sequence
- Shots as initial partition
- Similar shots are grouped in the same branches
- Provides an *Index* to the original sequence



Coding Scheme



1. Create / access BPT representation
2. Visit each node (pre-order traversal) to create new shuffled sequence
3. Encode normally the shuffled sequence
4. Decoder must arrange frames in visual order

Experimental Results (I)

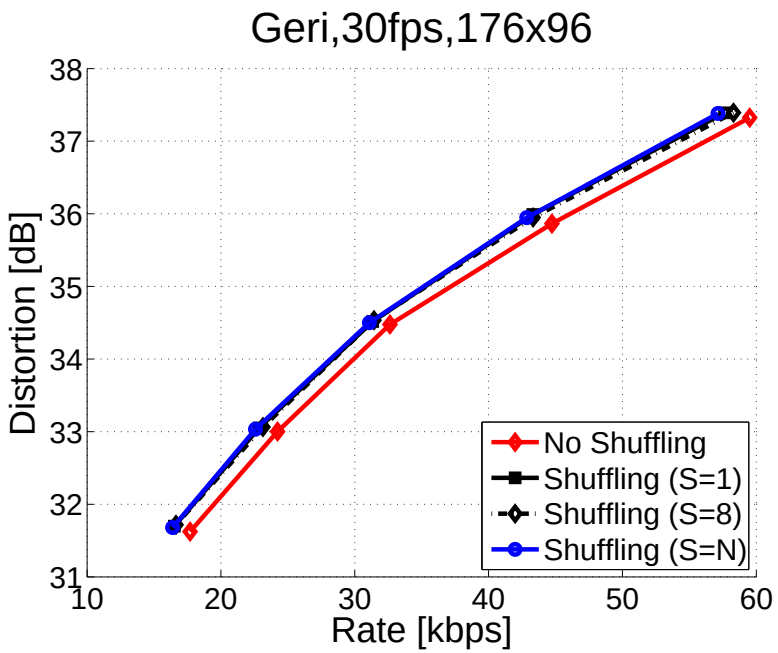
Original *Drama* sequence

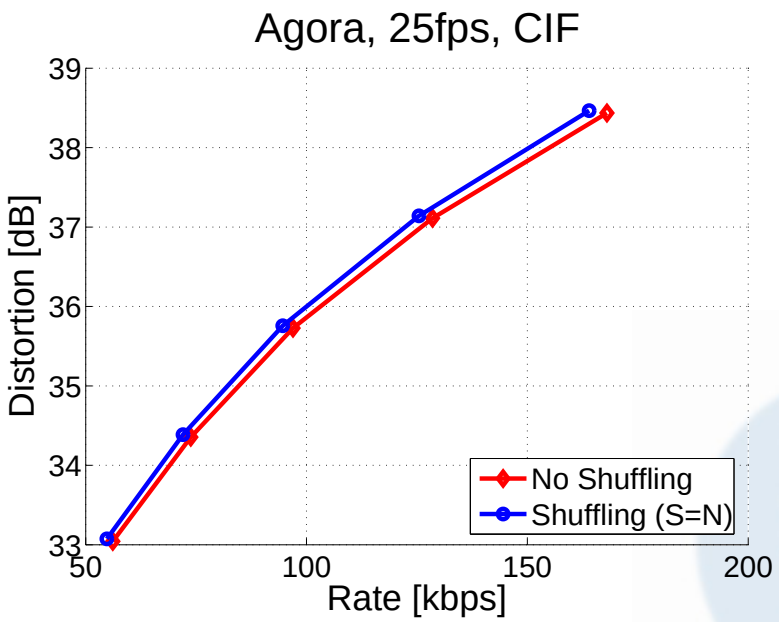
Shuffled *Drama* sequence

Experimental Results (II)

Indexing representation is available at the decoder

- Best results when using entire shots as initial partition
- Greater gains with sequences with scattered, short and similar shots





Experimental Results (III)

Indexing representation **not** available at the decoder

- List with video segment positions in the shuffled sequence:

$$\frac{2 \cdot S \cdot \lceil \log_2 F \rceil}{F}$$

- Indexing representation information is lost

Test	Data H.264 [kbps]	Savings [kbps]	Shuffling Information [kbps]
Geri (QCIF)	18.48	2.03 (11.0%)	0.012 (0.06%)
	58.82	2.79 (4.8%)	0.012 (0.02%)
Àgora (QCIF)	24.47	0.94 (3.8%)	0.002 (0.01%)
	91.44	2.86 (3.1%)	0.002 (0.00%)

Conclusions

Video Segment Shuffling using MPEG-7 Video Segment descriptors:

- ✓ Up to 10% bitrate savings at the same visual quality
- ✓ Greater gains with sequences with scattered, small and repeating shots
- × No gain for other sequences, but still interesting to store to easy access through a BPT
- × Only useful for non-real time applications

General Conclusions

Indexing representations:

- New indexing representation to summarize video shots
- Exploit compression representation by sharing mosaics
- Morphological operators to improve background information in mosaics
- Create key-regions that represent foreground objects

Compression representations:

- Four video coding techniques
- Exploit indexing representations extracted for other functionality
- Significant gains with respect to H.264

Future Work

Indexing representations:

- Integration of key-region in compression representations
- Activity of key-regions

Compression representations:

- Combination of compression schemes
- Improve existing techniques:
  - Long term reference *block* selection (ICASSP'07)
  - Other indexing representation (Edge Histogram, Color Structure, Homogeneous Texture)
- Other compression representations such as wavelets

Más Información

- Tesis doctoral
- Indexing representations:
  - J.Ruiz-Hidalgo and P.Salembier, "Robust segmentation and representation of foreground key-regions in video sequences", ICASSP'01.
- Compression Representations:
  - J.Ruiz-Hidalgo and P.Salembier, "On the use of indexing metadata to improve the efficiency of video compression", CSVT, 2006.
  - J.Ruiz-Hidalgo and P.Salembier, "Long term selection of reference frame sub-blocks using MPEG-7 metadata", ICASSP'07, Hawaii.
- *http://gps-tsc.upc.es/imatge/_Jrh*

Gracias!