samt 2006
1st International Conference on Semantic and Digital Media Technologies
former European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies
December 6-8, 2006 • Athens, Greece

Extraction of Motion Activity from Scalable-coded Video Sequences

Luis Herranz, Fabricio Tiburzi, Jesús Bescós

UAM UNIVERSIDAD AUTONOMA DE MADRID

Grupo de Tratamiento de Imágenes

1. Current Scene

Motion Activity

- Perceptual feature to describe the amount of action that shows a video segment
- Very useful for tasks as video Indexing and video analysis
- Widely addressed in last years and some approaches have been included into standards (MPEG descriptor)

Scalable Video

- One of the coding paradigms to address the current requirements of the evolving multimedia scenarios
- Sequences are coded in such a way that they can be very efficiently decoded at different fidelity levels.

Content Adaptation ←→Selection of certain parts of the adapted stream

However…

- Works that extend the existing approaches to calculate Motion Activity in Scalable Video are relatively scarce

MPEG-7 Motion Activity descriptor

- Calculated as a quantization of the standard deviation of the MPEG motion vectors
- Highly related with human perception of the “intensity of action”
- Defined specifically for MPEG coded sequences

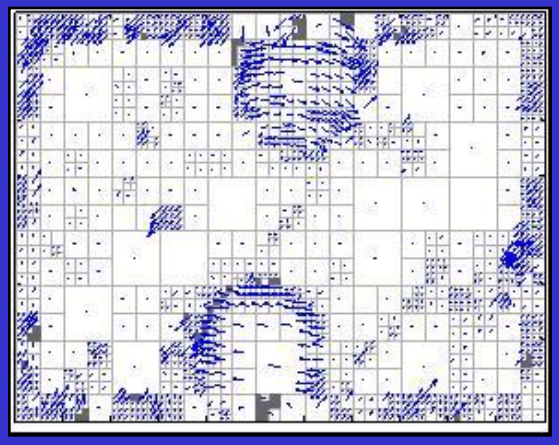
Goals of our work

- To extend the definition of the MPEG-7 Motion Activity descriptor to the context of **wavelet-based** scalable video
- To compare our extension with the MPEG case in terms of:
  - *Accuracy* : Evaluation of the descriptor behavior in a common application (Video Summarization)
  - *Efficiency*

2. Scalable codec overview description

- **Based on a t+2D wavelet framework**
  - It supports multiresolution in a natural way
  - It provides:
    - *Spatial scalability.* Through a 2D Discrete Wavelet Transform (2D DWT)
    - *Temporal scalability.* Through Motion Compensated Temporal Filtering (MCTF)

    = Discrete Wavelet Transform + Motion Compensation
    - *Quality scalability*
- **Motion compensation is performed using *Hierarchical Variable Size Block Matching* (HVSBM) and *forward* motion compensation**



- **Developed by QMUL. We will refer to this codec as "SVC"**

3. Extension of the motion activity descriptor for SVC

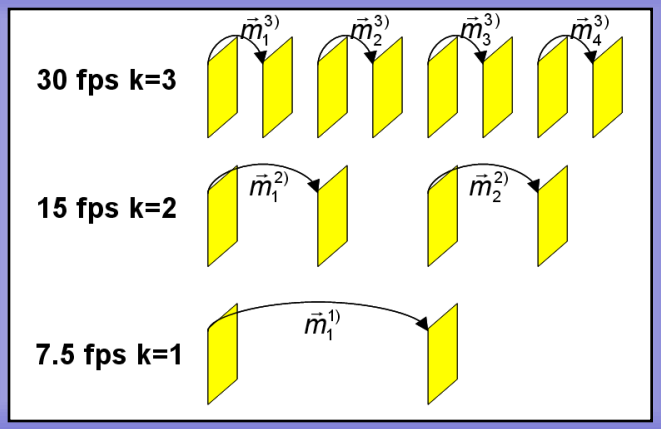
- MPEG-7 motion descriptor is defined for MPEG predicted frames and MPEG motion vectors
- In SVC there are also predicted frames and motion vectors but…

	Motion Vectors…	
	in SVC ▼	***in MPEG ▼***
1	…are intended to support the temporal scalability provided by the MCTF	…simply relate two consecutive frames
2	…represent displacement of different sized areas (HVSBM)	…represent displacement of uniform sized areas

- How 1 and 2 influence in the the Motion Activity computation in SVC?

3. 1 Dealing with (1) (MCTF)

- **For GOP of 8 frames and 3 levels of temporal scalability:**

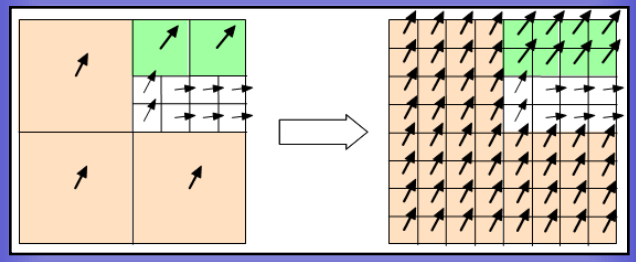


- **Conceptually in each level we have:**
  - The information that must be necessarily decoded to get the frames at that level
    - *For every odd frame: pixel information data (still in the 2D spatial wavelet domain)*
    - *For every even frame: motion vectors + residual coefficients*
  - The information that can be optionally decoded to increase the temporal resolution (ascend one level in the hierarchy)
    - *For every frame (even or odd): motion vectors + residuals coefficients of the following frame in the next level*

- **Therefore…**
  - A different set of motion vectors is needed to compensate motion in each level
  - If we compute the motion activity in each predicted-coded frame, its temporal resolution will depend on the temporal resolution of the working level.
- **It is necessary to speak about "Motion Activity at level k" (I^K)**
  - Its temporal resolution is one half of the temporal resolution of the level k.
  - How different is (and how much does it improve?) the motion activity in each temporal level? → We'll see it in the results section!

3. 2 Dealing with (2) (HVSBM)

- **Computation of the motion activity as the standard deviation of motion vectors *only makes sense if each vector refers to the motion of equally sized areas***
  - In MPEG 16 x 16 pixels areas (macroblocks)
- **It is straightforward to build an equivalent MPEG motion grid from a HVSBM motion grid**

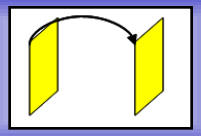


- **This replication can be efficiently done by weighting each vector by the number of 16x16 areas that it includes.**
  - The bigger the areas covered by the motion vectors in HVSBM, the more computation efficiency increases with respect to the MPEG case

4. Motion Activity evaluation procedure

- **By contrast with the MPEG-7 Motion Activity through a common application of the feature: video summarization (by key frame selection)**

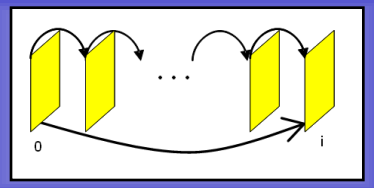
4. 1 Overview of the key frame selection scheme



Motion activity of one frame

STRONGLY RELATED WITH…

Amount of information added by that frame respect to the previous one



Cumulated motion activity of a frame i through a continuous flow of prediction from frame 0.

STRONGLY RELATED WITH…

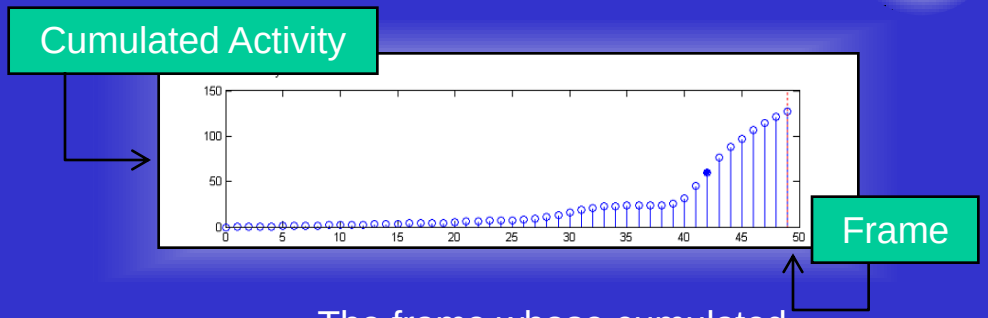
Amount of information added by frame i respect to frame 0

Most likely desired Key-Frame of a sequence

Not first frame
(Nothing of information added)

Not last frame
(All the information added)

The frame that has added the half of the total information

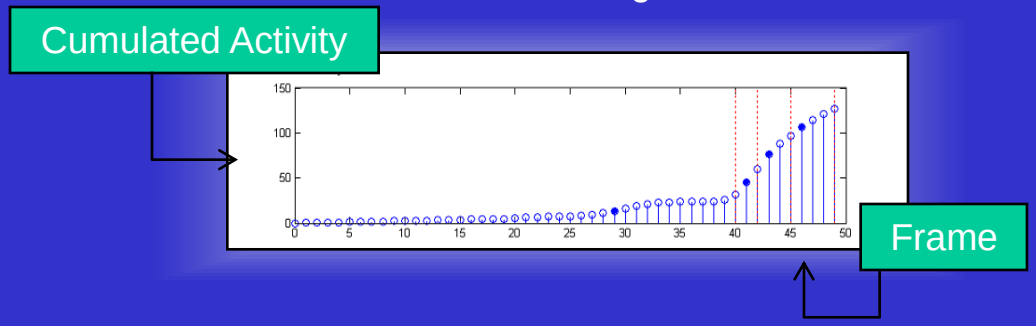


The frame that have added the half of the total information

STRONGLY RELATED WITH…

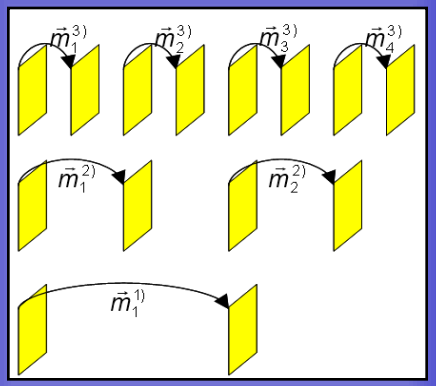
The frame whose cumulated motion activity is half of the maximum

- **Therefore, to select n key-frames**
  - We divide the sequence in n equal segments in the cumulated motion activity scale
  - We select the frames at the middle of each segment.



- **In this way all the key frames represent theoretically the same amount of information of the sequence.**

- **Little problem in SVC: There isn't an interframe continuous prediction in any level (like in MPEG excepting I frames).**
  - Therefore there isn't an activity value for each frame→ the accumulation of the activity is only "partial"



- **The frames (or sets of frames) that are not predictively related at level k, are predictively related at levels k-1 or k-2...or 1**
- **We add to our study a new level of activity: *"inter-level activity"* in which the activities obtained in every temporal level are summed to reflect the previous fact**

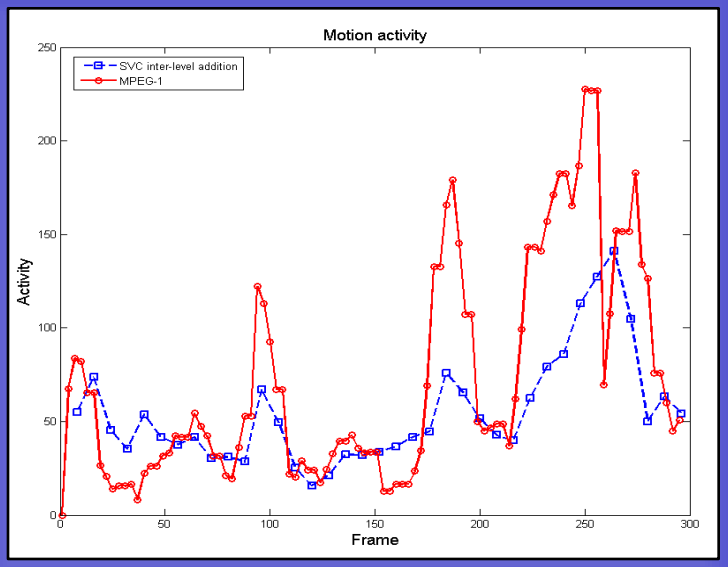
4. 2 Comparison of SVC and MPEG obtained key-frame sequences

- **Performed by contrasting the quality of the final summaries**
- **Several deterministic measures have been proved to be highly correlated with subjective evaluation:**
  - SemiHausdorff distance: It measures the fidelity of a **SET** of key-frames respect to a **SET** with the frames of the sequence.
    - Good measure if the temporal structure of the sequence has not been taken into account in the summarization task (ex. summarization by clustering)
  - Distortion distances: They measure the fidelity of a **SEQUENCE** of (maybe replicated) key-frames respect to the original sequence.
    - Good measure if the temporal structure of the sequence has been considered in the summarization task
- **In our key-frame selection scheme the position of every frame in the sequence is essential → We use a distortion metric.**
  - This metric computes the frame distance in the Principal Component space

5. Results and conclusions

- **We have tested our approach in the *Stefan* and *Foreman* sequences**
  - MPEG coded with:
    - GOP of 15 frames IBBPBBPBB…
    - 30 fps
  - SVC coded with:
    - GOP de 8 frames
    - 3 temporal descompositions (30 fps, 15 fps, 7.5 fps)
    - 3 spatial descompositions (352x288, 176x144, 88x72)
    - HVSBM with minimum block size =8x8 and maximum block size=64x64

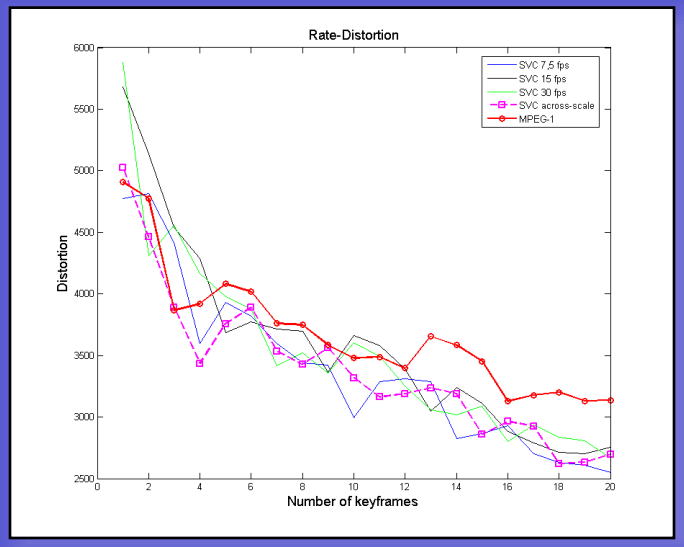
Extracted motion activity (Stefan sequence):



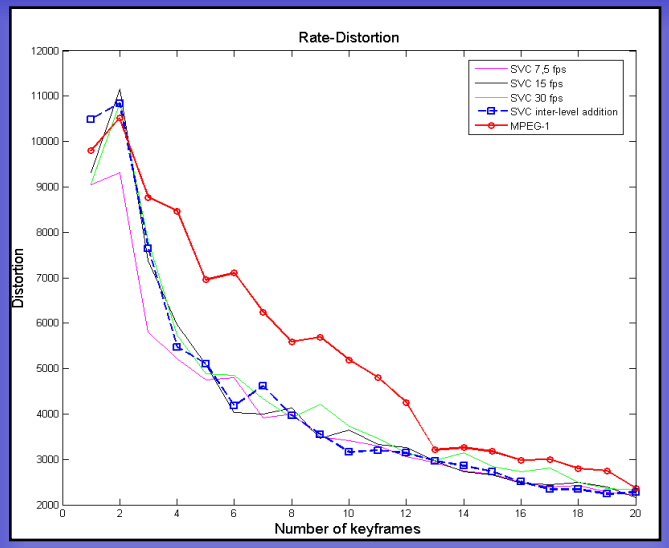
- SVC curve obtained combining the three temporal levels (but the results of each level are practically identical to this one)
- Both curves follow approximately the activity present in the sequence
- The range of variation is wider in the case of MPEG

Rate(number of frames)-distortion curves

Stefan



Foreman



- MPEG and SVC distortion curves follow a similar shape → THIS VALIDATE OUR APPROACH
- MPEG curves show slightly more distortion than those obtained for SVC
- The summary using the inter-level measure gives the most smooth result, but it almost doesn't improve the distortion of the rest of levels. Choice of a level should depend on other considerations (efficiency, availability of motion vectors…)
- Working on the lowest temporal level can be enough for the most of the applications, as the quality of the results its similar to higher resolutions

Questions?