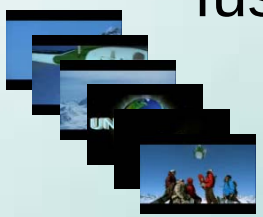
MUSCLE Showcase:

Movie Summarization and Skimming Demonstrator

ICCS-NTUA (P. Maragos, K. Rapantzikos, G. Evangelopoulos, I. Avrithis)
AUTH (C. Kotropoulos, P. Antonopoulos, V. Moschou, N. Nikolaidis, I. Pitas)
INRIA-IRISA (P. Gros)
TSI-TUC (A. Potamianos, M. Perakakis)

Audio-Visual Attention Modeling – Event Detection

- Detecting events by attention modeling
- Two-module (aural, visual) attention for 3D event histories
- Attention curve extraction. Fusing streams vs. fusing features



Audio Saliency

- Audio signal model:

sum of AM-FM components

$$s(n) = \sum_{k=1}^{K} A_k(n) \cos[\Phi_\kappa(n)]$$

- Modulation bands through a linear bank of *K* Gabor filters.
- Tracking the *maximum average Teager Energy* (MTE)

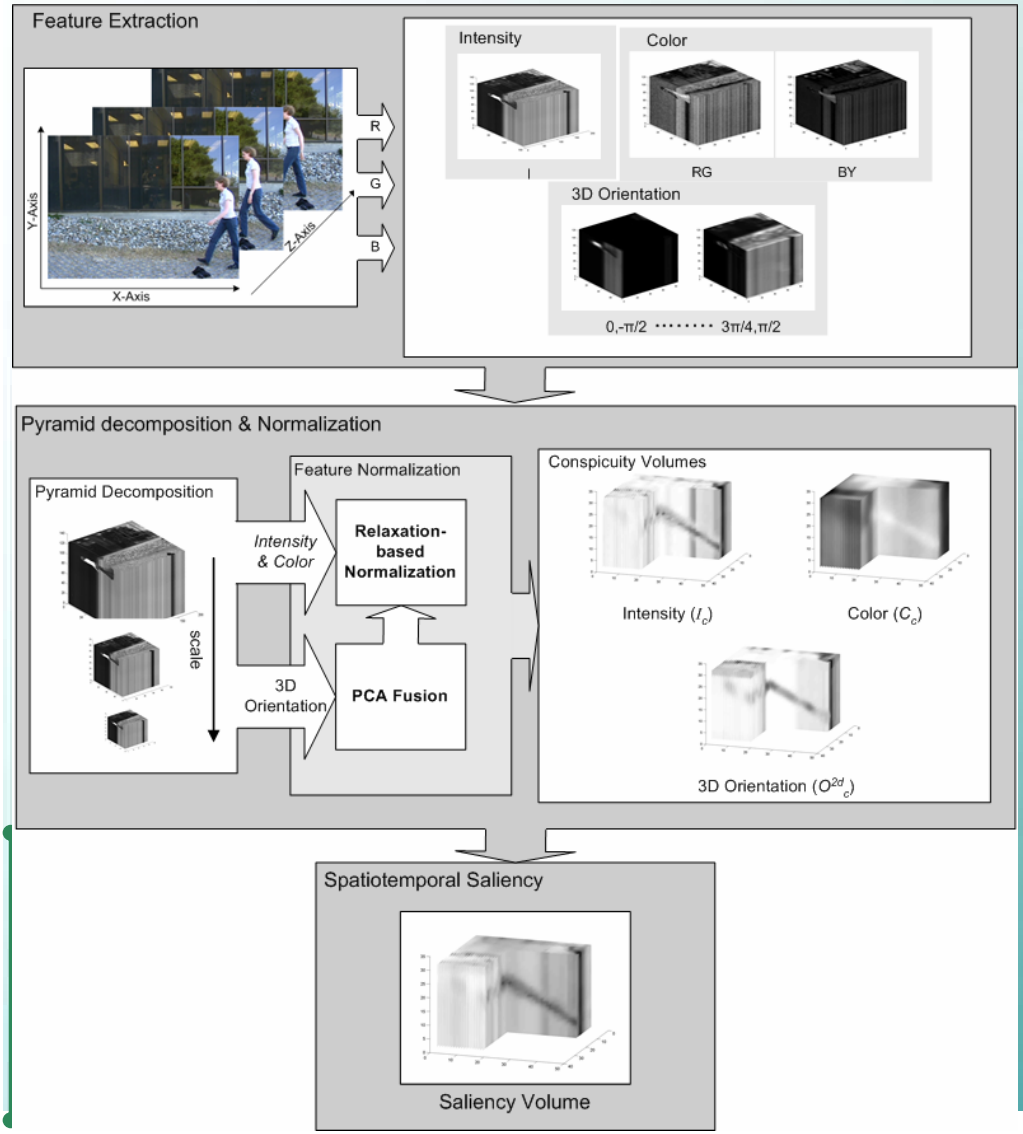
$$MTE(m) = \max_{1 \le k \le K} \frac{1}{N} \sum_{n=1}^{N} \Psi\left[\left(s * h_k\right)(n)\right]$$

- h_k : k-th filter response, Ψ :Teager-Kaiser Energy operator
- MTE : *dominant* signal *modulation energy*.
- Demodulating, via DESA, the dominant channel and frame average

$$MIA(m) = \frac{1}{N} \sum_{n=1}^{N} \left| A_i(n) \right|$$

$$MIF(m) = \frac{1}{N} \sum_{n=1}^{N} \left| \Omega_i(n) \right|$$

Spatiotemporal Visual Saliency



Features

- Intensity
- Color
- Spatiotemporal orientations

Feature intra- and inter- competition

$$\frac{\partial E}{\partial F^k(c)} = \lambda_D \cdot \frac{\partial E_D}{\partial F^k(c)} + \lambda_S \cdot \frac{\partial E_S}{\partial F^k(c)} =$$

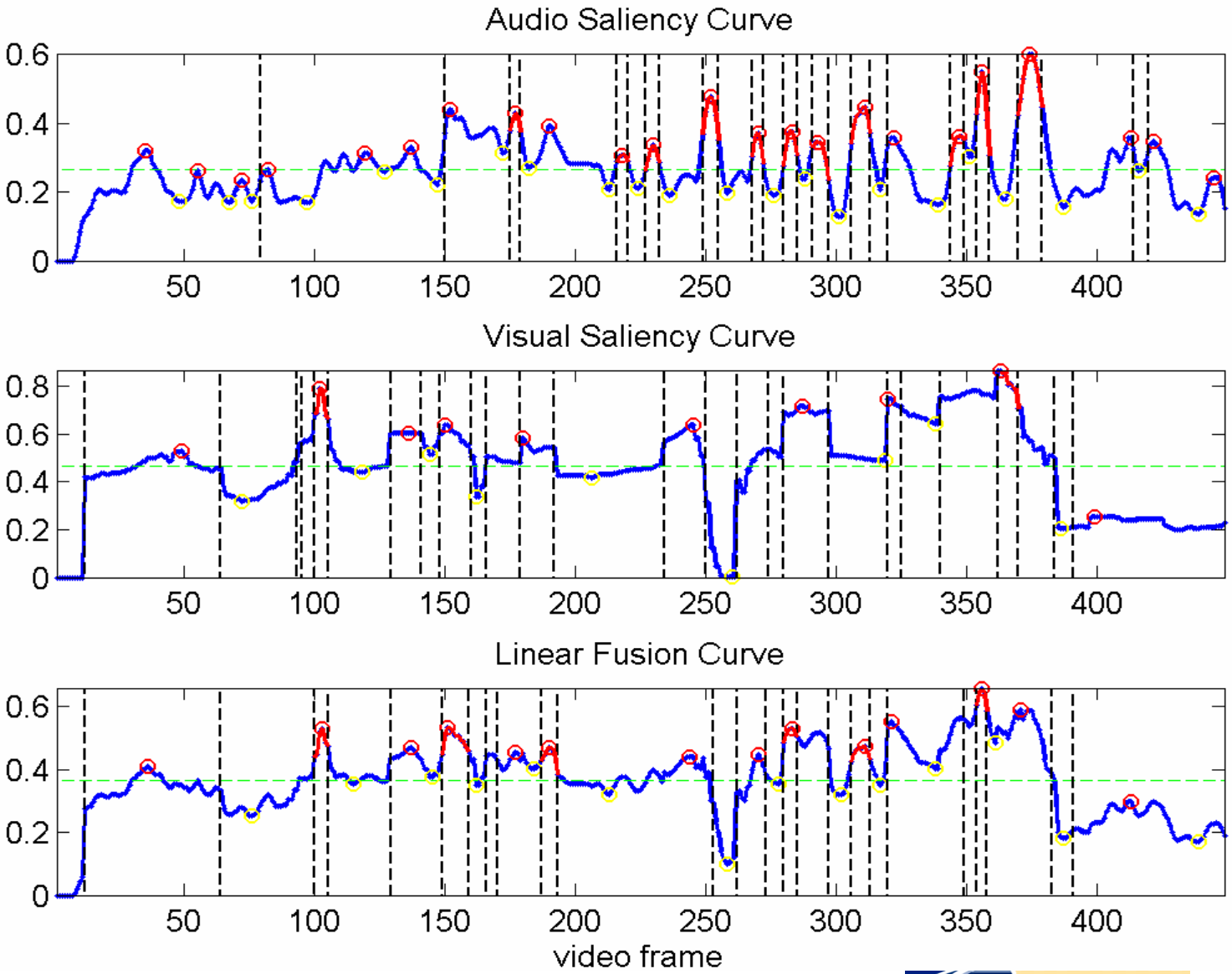
$$= \lambda_D \cdot \left(\left| F^k(c) - F^k(h) \right| + F^k(c) \right) +$$

$$+ \lambda_S \cdot \frac{1}{card(Q)} \cdot \left(\sum_{q \in Q} F_q^k(c) + \sum_{q \in Q} O_c^{3D} \right)$$

AudioVisual Fusion – User attention curve

- Simple linear fusion scheme $M = \vec{w}_v \cdot \vec{V} + \vec{w}_a \cdot \vec{A}$
- Detecting events by 4 curve characteristics:
  - *Peak/valley* detection (key-frame selection)
    - Local maxima\minima
  - Sharp transition detection (1D *edges*)
    - LoG operator on curve
    - Scale parameter by std of Gaussian
  - *Thresholding* values (salient segments)
  - Region of peak support (lobes, segments between edges where maxima exist)
- Two fusion schemes:
  - i) Fuse curves (linear, non-linear fusion)
  - ii) Detect in audio and video and combine (e.g. AND,OR)

User Attention Curve

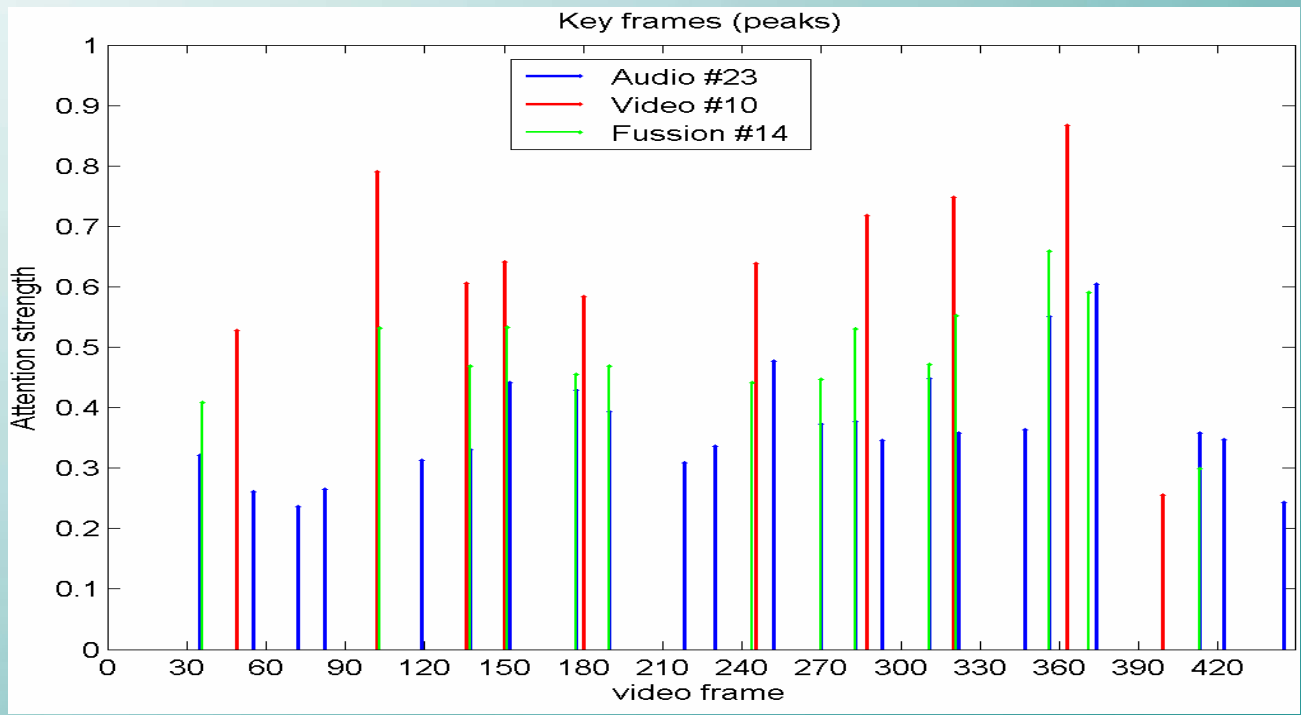


Key frame selection

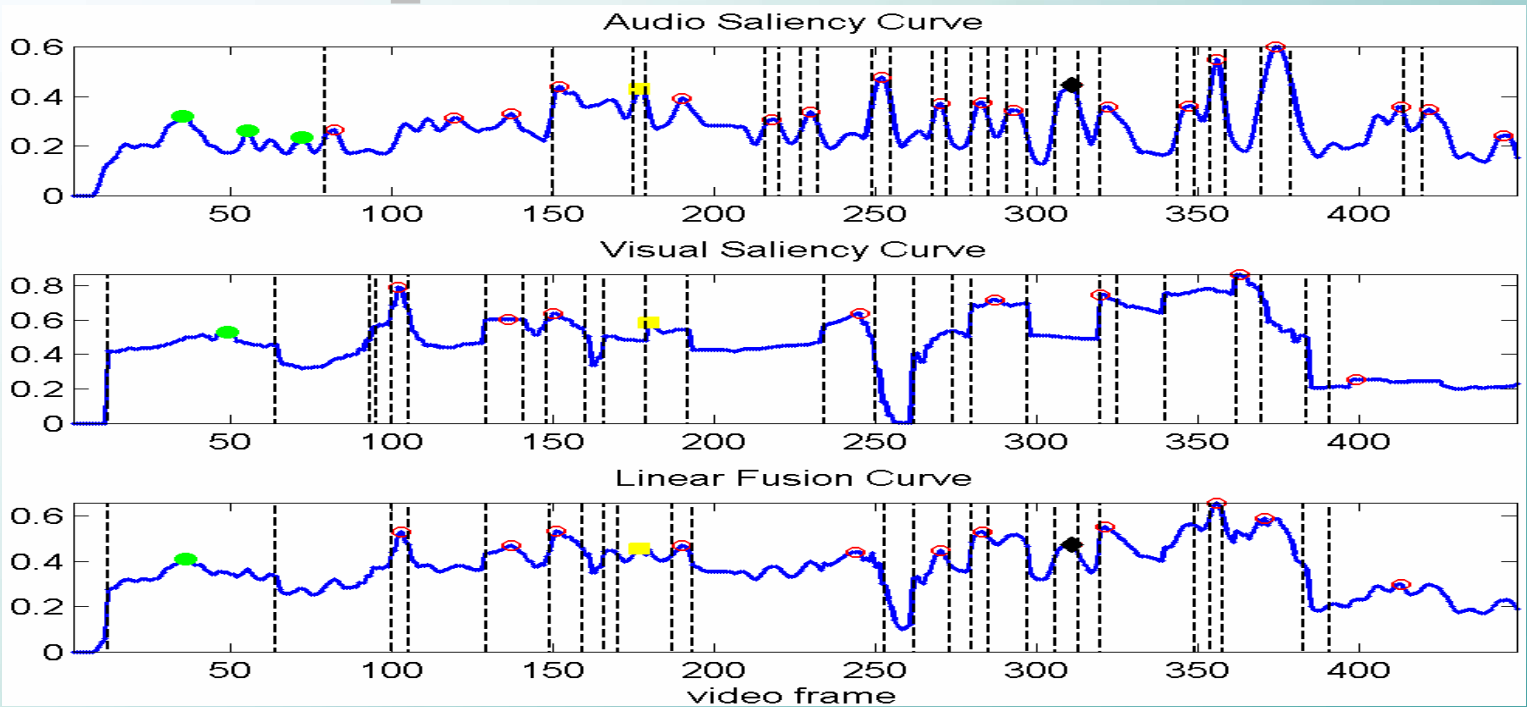
Audio

Video

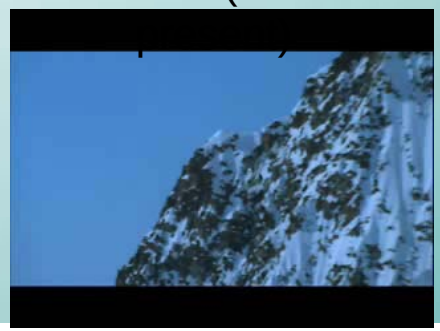
Fusion



Examples of Audio/Video



- Video suppresses/groups audio events (audio event



- Audio & Video events match (both are present)

- Audio giving event (video event absent)

Movie Database Description

- 42 scenes were extracted from 6 movies of different genres, i.e., Analyze That, Lord of the Rings, Secret Window, Platoon, Jackie Brown, Cold Mountain.
- 25 out of the 42 scenes are dialogue instances and the remaining 17 are annotated as non-dialogue scenes.
- Dialogue scenes last from 20 sec to 120 sec.
- Total duration: 34 min and 43 sec.

Scene Annotation

- **Dialogue types** for both audio and video streams are:
  - CD (Clean Dialogue)
  - BD (Dialogue with background)

- **Non-Dialogue** types for both audio and video streams are:
  - CM (Clean Monologue)
  - BM (Monologue with background)
  - ND (Other)

Database Description

- *gt folder*: ground truth information (*.xml files).
- *video folder:* the video streams without the audio channel (*.avi files).
- *audio folder*: the audio streams without the visual channel (*.wav files).
- *actors index*: actor's Id, name, and photograph (*.xls file).
  - Actors info is also available in xml format for each video scene.