MUSCLE Showcase:

Movie Summarization and Skimming Demonstrator

ICCS-NTUA (P. Maragos, K. Rapantzikos, G. Evangelopoulos, I. Avrithis, S. Kollias)

AUTH (C. Kotropoulos, P. Antonopoulos, V. Moschou, N. Nikolaidis, I. Pitas)

INRIA-Texmex (P. Gros, X. naturel)

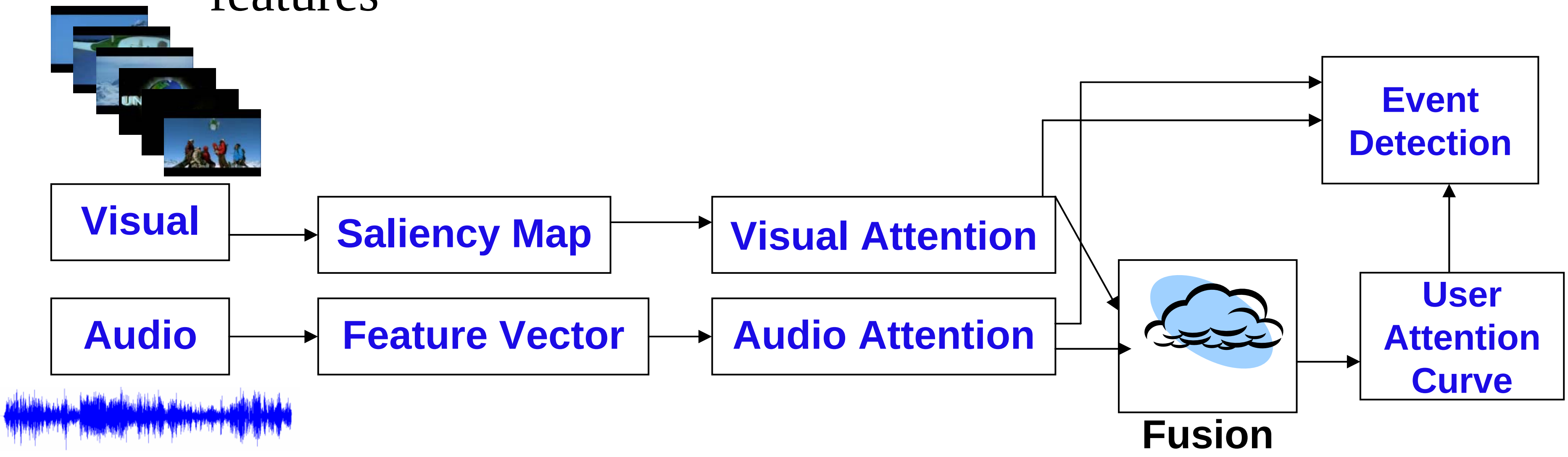
TSI-TUC (A. Potamianos, M. Perakakis)

Partners

- **ICCS-NTUA (leader)**
  - □ Design and develop AudioVisual Saliency estimators. Abrupt-change Detectors. Pre-segmentation around key frames.
- **AUTH**
  - □ Provide a movie database along with appropriate annotation. Collaborate on AV Saliency detection.
- **INRIA-Texmex**
  - □ Statistical models for video/scene segmentation.
- **TUC**
  - □ Design and implement the user interface

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 Modeling and Features

- 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)

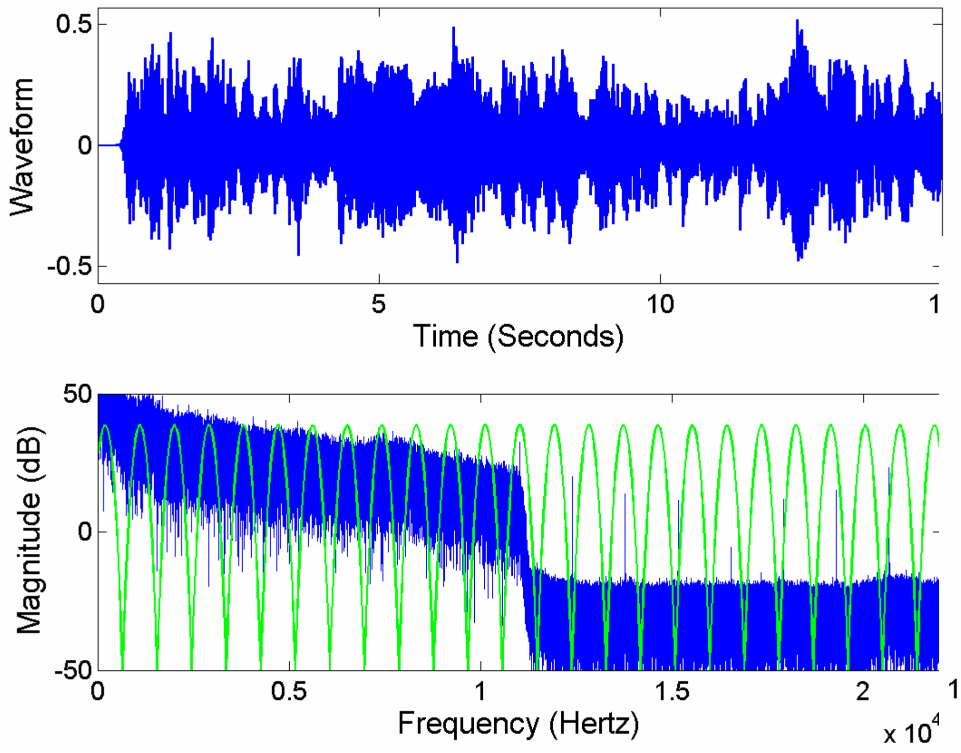
$$MTE(m) = \max_{1 \le k \le K} \frac{1}{N} \sum_{n=1}^{N} \Psi\left[\left(s * h_k\right)(n)\right]$$

- 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

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

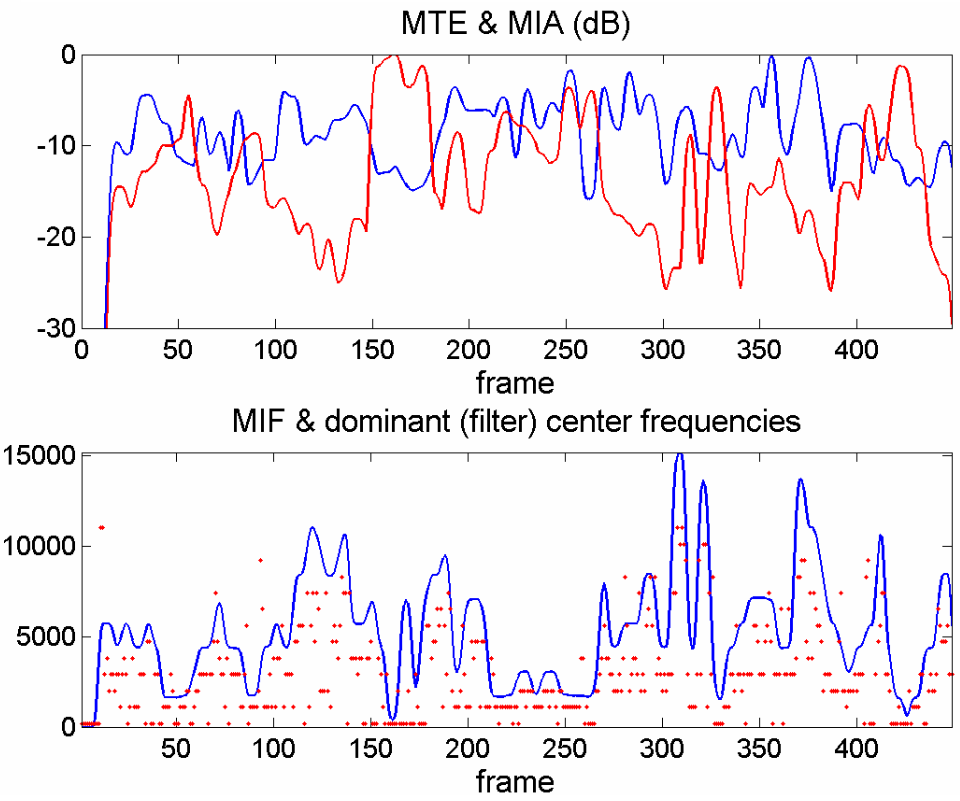
$$MIF(m) = \frac{1}{N} \sum_{n=1}^{N} \left|\Omega_i(n)\right|$$

Feature Vector Formation



3D normalized feature vector

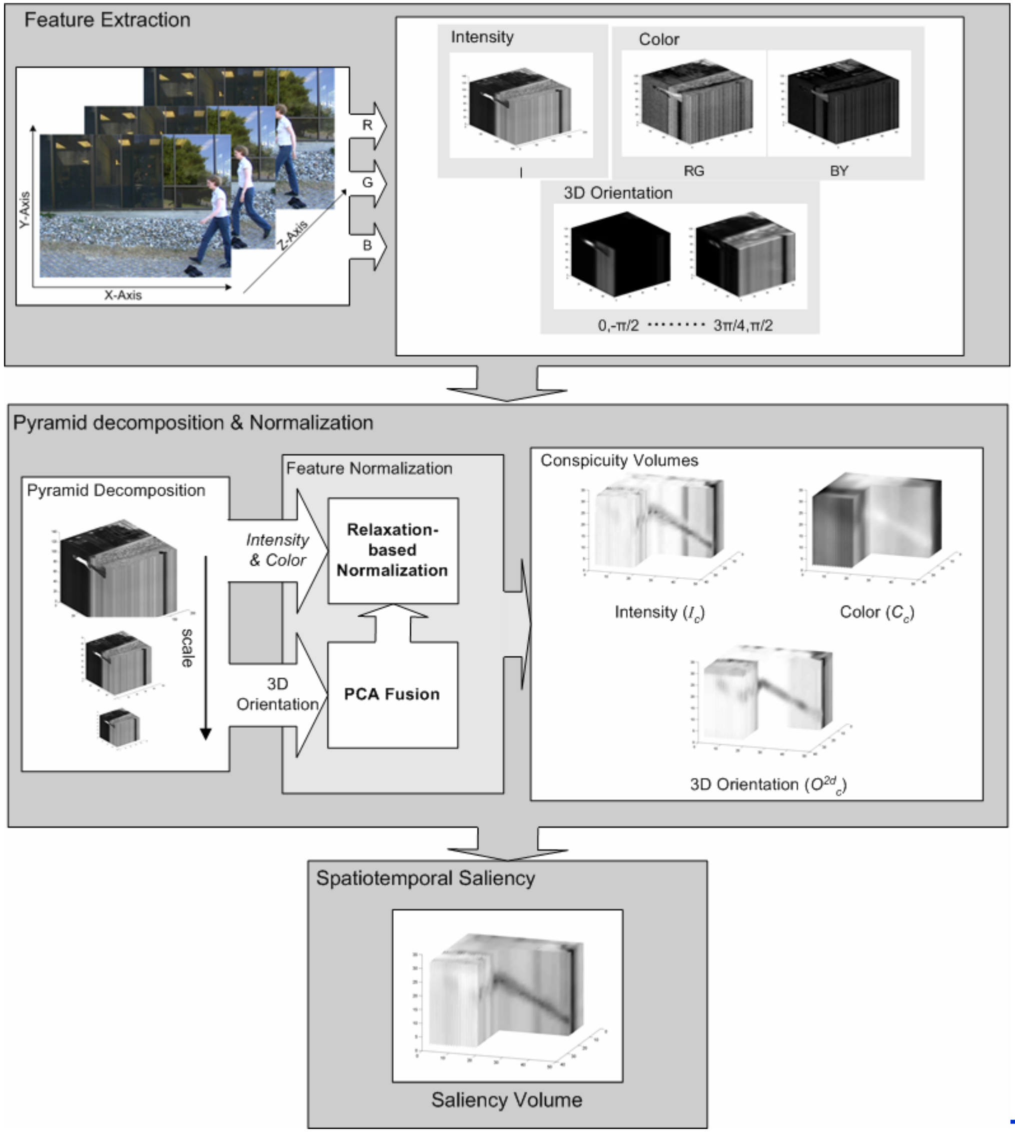
$$\boxed{\vec{A} = \{A_i\} = \{MTE, MIA, MIF\}}$$


MTE & MIA (dB)

MIF & dominant (filter) center frequencies

- ❑ Audio window to video frame index map (e.g. decimation, max)

Spatiotemporal Visual Saliency



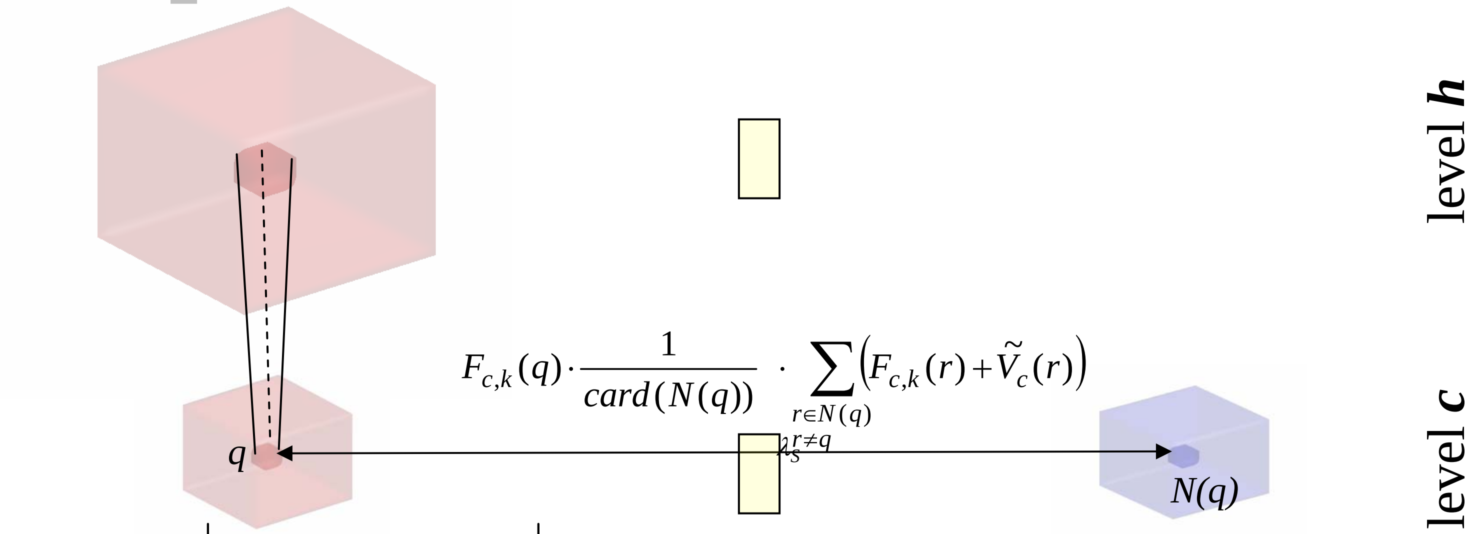
Features *(F)*

- Intensity *(I)*
- Color *(RG, BY)*
- Spatiotemporal orientations ($\tilde{V}$)

Steps

- Pyramidal decomposition
- Normalization & Fusion
- Conspicuity volumes generation
- Saliency volume computation

Visual Saliency model: Feature Competition



$$F_{c,k}(q) \cdot \frac{1}{card(N(q))} \cdot \sum_{\substack{r\in N(q)\\ r\neq q}} \left(F_{c,k}(r) + \tilde{V}_c(r)\right)$$

$$F_{c,k}(q) \cdot \left|F_{c,k}(q) - F_{h,k}(q)\right|$$

Motion activity

Iterative energy minimization scheme that acts on 3D local regions and is based on center-surround inhibition constrained by inter- and intra- local feature values.

$$\frac{\partial E}{\partial F_{c,k}(q)} = \lambda_D \cdot \frac{\partial E_D}{\partial F_{c,k}(q)} + \lambda_S \cdot \frac{\partial E_S}{\partial F_{c,k}(q)} =$$

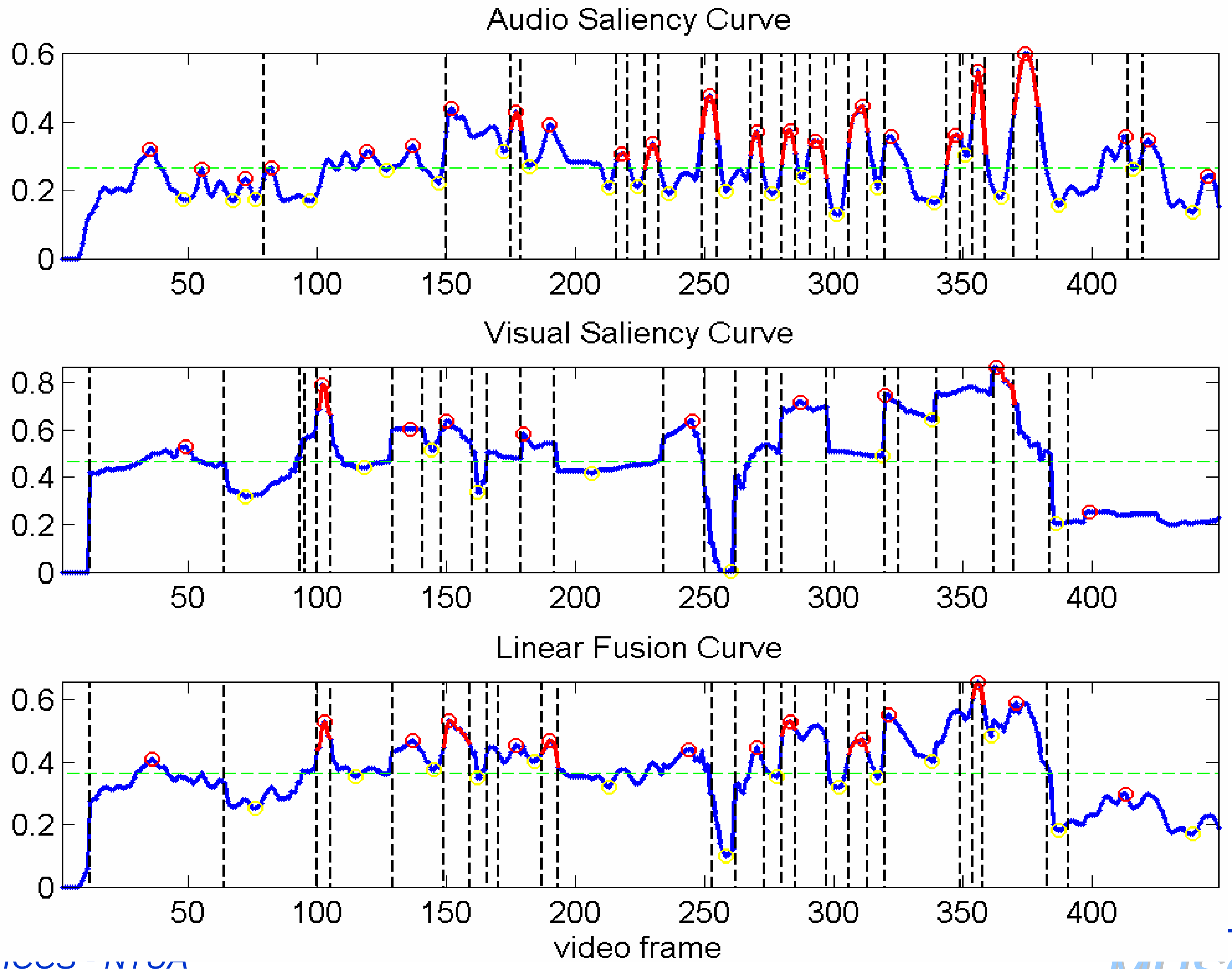
$$= \lambda_D \cdot \left(\left|F_{c,k}(q) - F_{h,k}(q)\right| + sign\,(F_{c,k}(q)) \cdot F_{c,k}(q)\right) + \lambda_S \cdot \frac{1}{card\ (N(q))} \cdot \sum_{\substack{r\in N(q)\\ r\neq q}} \left(F_{c,k}(r) + \tilde{V}_c(r)\right)$$

$$F = \{\mathrm{I}, \mathrm{RG}, \mathrm{BY}\}, \quad k \in \{1, \ldots, card(F)\}$$

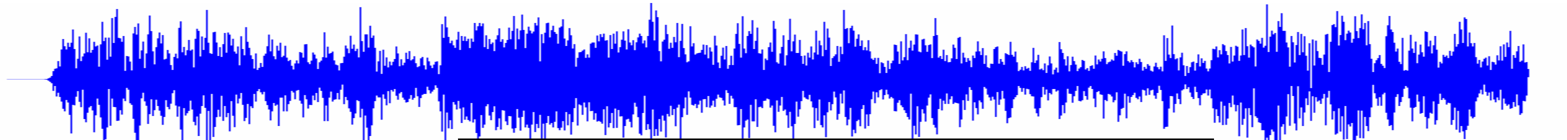
AudioVisual Fusion – User attention curve

- 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)

Saliency Curves



Example (Movie trailer)



www.firstdescentmovie.com

- Movie trailer (mpeg): 15sec, 30frames/sec
- Rich in Events:
  - Visual (color, motion, action shots, persons, objects, text)
  - Audio (helicopters, noises, music, speakers, transmissions, effects)

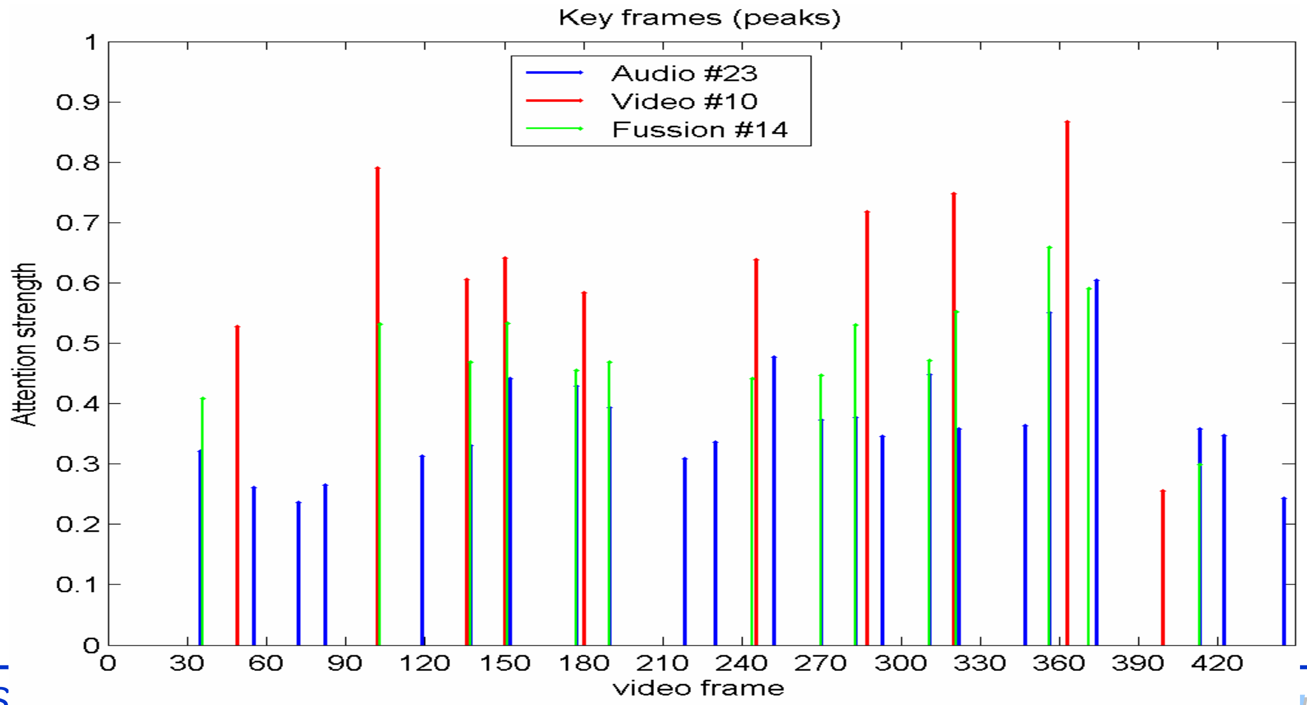
Event detection based on peaks (fusion curve)

Key frame selection

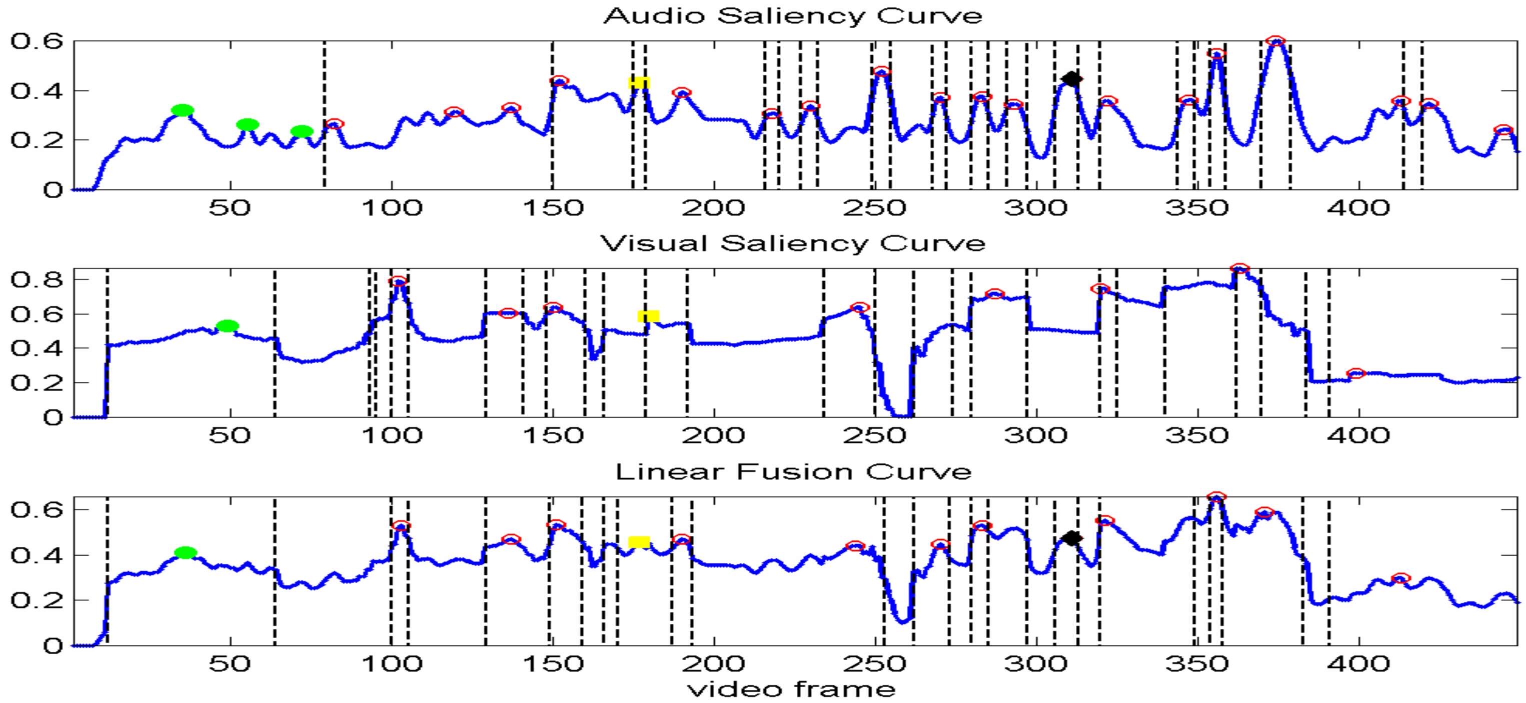
Audio

Video

Fusion



Examples of Event Detection

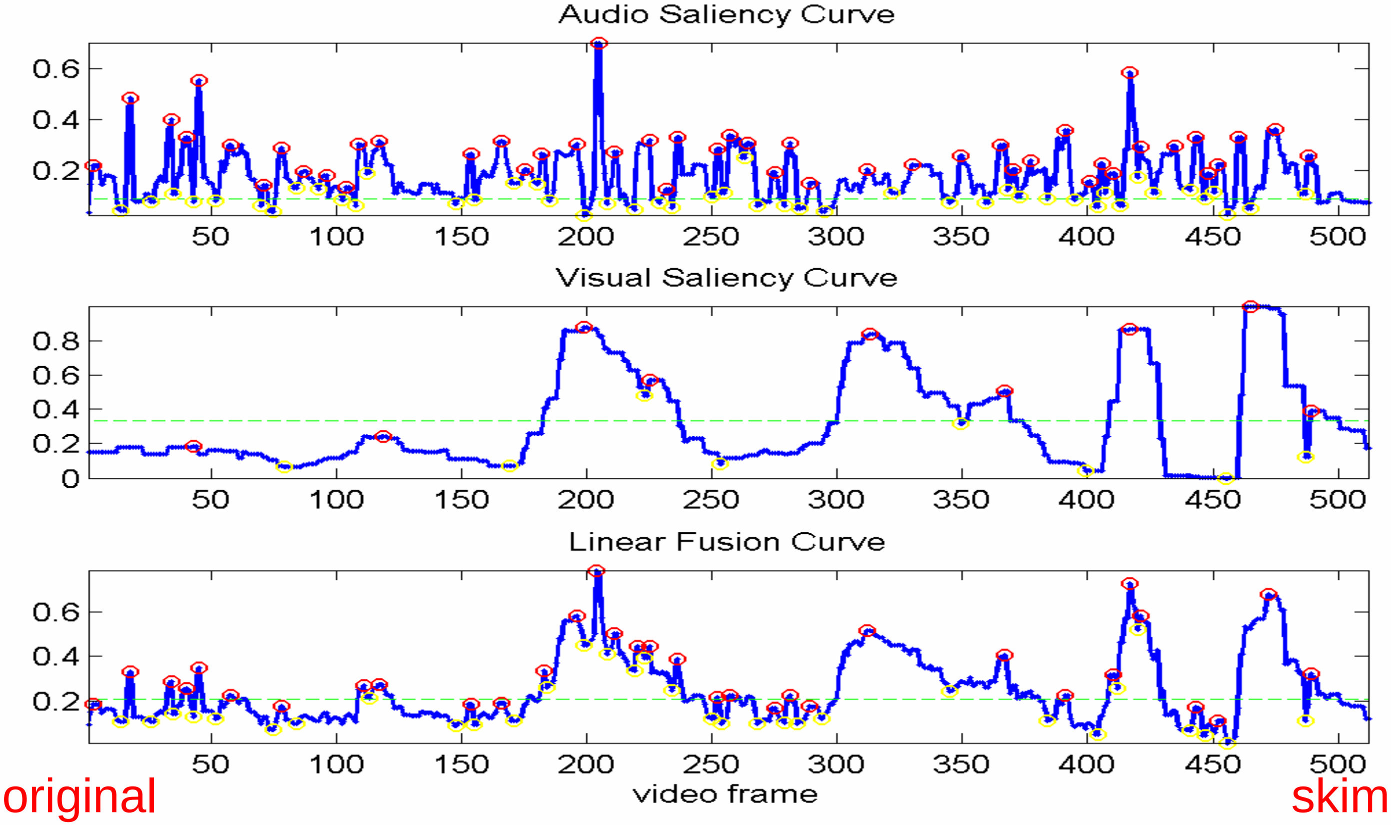


● Video suppresses/groups audio events (audio event present)

🟨 Audio & Video events match (both are present)

■ Audio giving event (video event absent)

Examples of Event Detection: AUTH database



Audio Saliency Curve

Visual Saliency Curve

Linear Fusion Curve

original

skimmed

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.

Current Scene Annotation

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

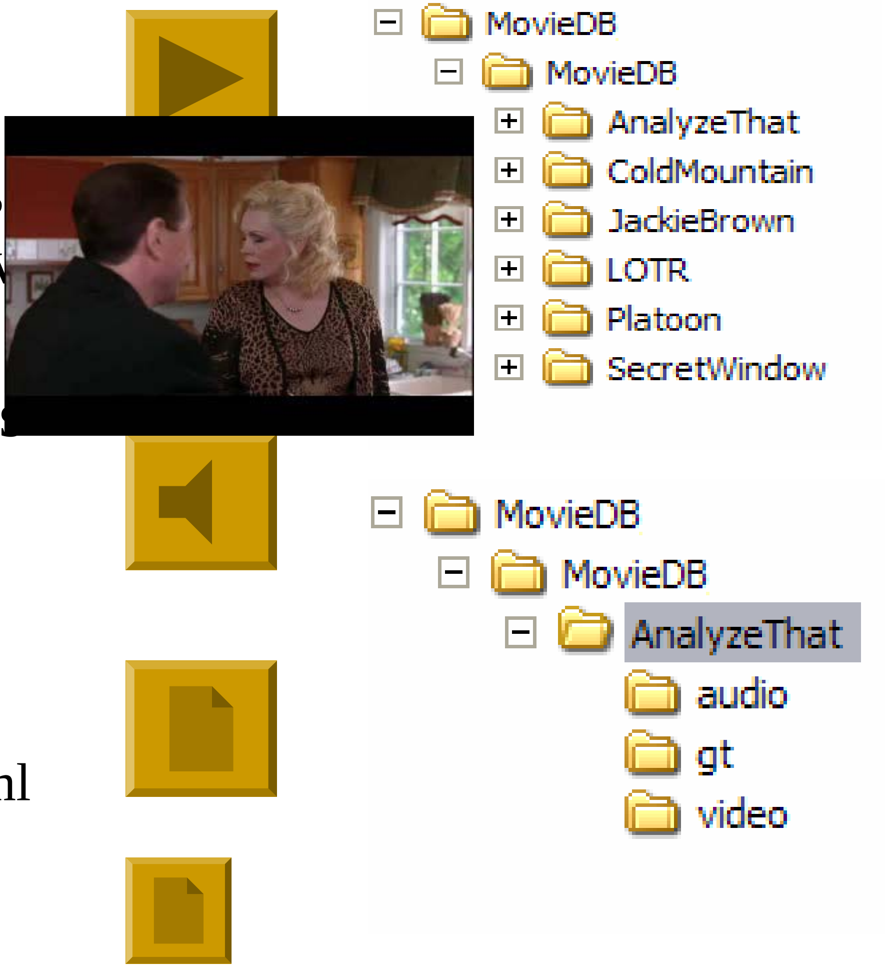
- **Non-Dialogue** types for both audio and video streams are:
  - ❑ CM (Clean Monologue)
  - ❑ BM (Monologue with background)
  - ❑ ND (Other)

Extended Scene Annotation

- **Motivation**
  - The notion of saliency is quite subjective
  - Human evaluation needed to ensure “objectivity”
- **Objective**
  - Create annotation useful for evaluating saliency detection methods
- **Use 3 levels of annotation**
  - Audio only
  - Visual only
  - Audiovisual

Database Description

- *gt folder*: ground truth information (*.xml files).
- *video folder:* the video streams without the audio channel (*.av 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.



- ⊟ MovieDB
  - ⊟ MovieDB
    - ⊞ AnalyzeThat
    - ⊞ ColdMountain
    - ⊞ JackieBrown
    - ⊞ LOTR
    - ⊞ Platoon
    - ⊞ SecretWindow

- ⊟ MovieDB
  - ⊟ MovieDB
    - ⊟ AnalyzeThat
      - audio
      - gt
      - video

Selection and Learning of Salient Events (INRIA)

- Generic solution of selection (1)
  - Select a subset of salient events: global minimization of redundancy between salient events
- User-oriented solution
  - Goal: provide a summary based on user specifications
  - Learn parameters of user-specified events
  - Select salient events according to the learning phase and method (1)

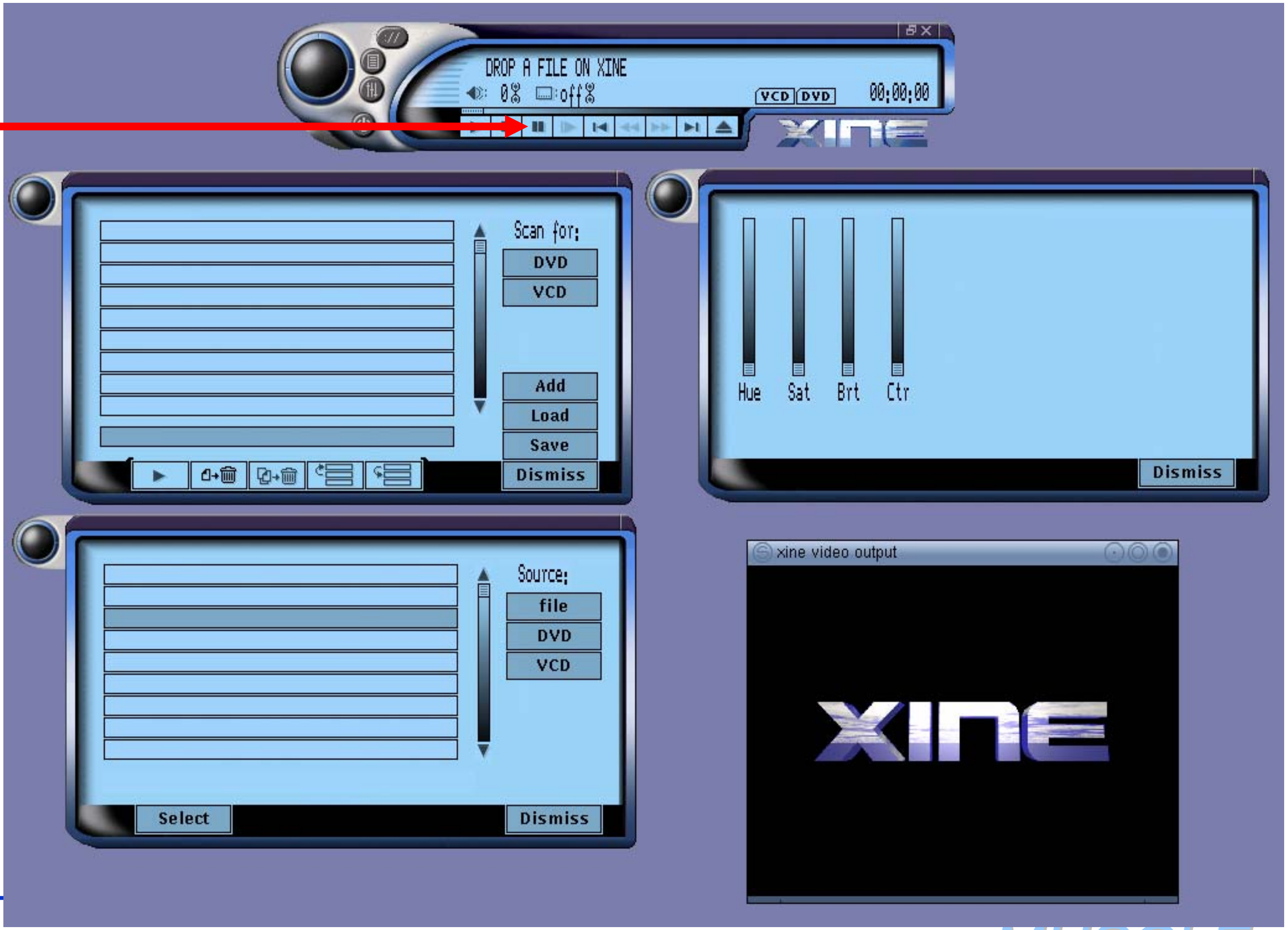
Movie Summarizer Player UI (TUC)

- User selects the degree of summarization
  - Available levels: none, ½, ¼, trailer
- User can change the level at any time
- System pre-renders the movies at the four levels of summarization
- Movie player based on xine open-source multimedia player
- xine: written in C++, easy to modify, lost of features, light version also available

Example xine player control

Add summarization level control buttons

x2 x4 xM



Current Status & Future Work

- **Current Status**
  - **Baseline version is available**
    - **Audio saliency module**
    - **Video saliency module**
    - **Simple audiovisual fusion approaches have been adopted**
    - **Experiments on the AUTH database have been undertaken**

- **Next steps…**
  - **Extension of AUTH database annotation**
  - **Statistical models for audiovisual segmentation**
  - **Design & implementation of a user friendly interface**