1. Outline
- Attention modeling of audiovisual streams.
- Saliency measures:
  - Visual: spatiotemporal attention model driven by intensity, colour and motion.
  - Audio: signal modulation and multifrequency band features (nonlinear operators and energy tracking).
- Fusion on a single audiovisual measure: Capture events across modalities by curve geometric features.
- Application to video abstraction: skimming and summarization.
- Saliency-based movie summarization algorithm.
- Bottom-up, content-independent, generic.
- Subjective evaluations: informativeness and enjoyability of different-gene movie summaries.

2. System overview

![Image 1](movie_summarization.png)

Figure 1. Overview of the system.

![Image 2](audio_visual_saliency.png)

Figure 2. 50% saliency-based real audio summarization.

![Image 3](audio_analysis.png)

Figure 3. Movie frames and corresponding spatio-temporal saliency volume slices.

![Image 4](audio_visual_fusion.png)

Figure 4. Key-frames (main noted by circles and shown) A human vs. automatic saliency annotations for medium- filtered AVS curve (300 movie, 800 frames).

![Image 5](summarization_algorithm.png)

Figure 5. Saliency-based summarization and construction of video skims.

3. Audio Analysis
- Audio AM-FM model
  \[ |e(t)|^2 = \sum_{l=1}^{L} \psi_l^2 |A_l(t)|^2 \phi_l(t) \]
- Gabor filterbank & demodulation
- Dominant modulation energy

![Image 6](audio_visual_salient_event_detection.png)

Figure 6. Best salient key-frames from "500" movie clip (approx. 7min).

![Image 7](subjective_evaluations.png)

Figure 7. Subjective evaluation scores of video skims (3 movie scenes) in different rates (2, 2.5, 3 real time), using audiovisual (AV) and multimodal (AVT) saliency.

4. Audio Visual Analysis
- Intra- and inter-frame competition

![Image 8](subjective_evaluations.png)

Figure 8. Subjective evaluation scores of video skims (3 movie scenes) in different rates (2, 2.5, 3 real time), using audiovisual (AV) and multimodal (AVT) saliency.

5. Audiovisual Fusion
- Audiovisual fusion

![Image 9](audio_visual_salient_event_detection.png)

Figure 9. Most salient key-frames from "500" movie clip (approx. 7min).

![Image 10](subjective_evaluations.png)

Figure 10. Subjective evaluation scores of video skims (3 movie scenes) in different rates (2, 2.5, 3 real time), using audiovisual (AV) and multimodal (AVT) saliency.

6. Summarization Algorithm
1. Filter: AVSC with radius of length 2M + 1.
2. Threshold choice: S_white = c \cdot \text{length (AVSC)}, D = a \cdot S_black > S_black
3. Selection: segments S_white > S_black
4. Reject: segments shorter than N frames $\rightarrow [l_{k-1}, l_k] = \emptyset$
5. Join: segments less than N frames apart $\rightarrow l_k = l_{k-1}$
6. Render: Linear overlap-aided on 2 video frames and audio.

![Image 11](subjective_evaluations.png)

Figure 11. Subjective evaluation scores of video skims (3 movie scenes) in different rates (2, 2.5, 3 real time), using audiovisual (AV) and multimodal (AVT) saliency.

7. Evaluation
- 3 clips from MUSCLE movie database, scenes (5-7 min) from "Lord of the Rings F" (LOTRI), "300" and "Cold Mountain" (CM).
- Skims for c = 0.5, 0.1, 0.2 (x2, x3, x5 real time).
- Correspondence with summarization saliency.
- Subjective evaluation: 11 naive users rated original & skims w.r.t. included information and aesthetics.
- Extension: Text info from subtitles (AVT saliency).

8. Relevant work

9. Acknowledgments
This work was supported by the European Union IST-PROTOS Network of Excellence ‘MUSCLE’, and in part by the EU research program ‘HIWIRE’. We would like to thank C. Kotropoulos and his group at Aristotle University of Thessaloniki (AUTH) for providing the movie database.