Computer Vision, Speech Communication & Signal Processing Group,
Intelligent Robotics and Automation Laboratory
National Technical University of Athens, Greece (NTUA)
Robot Perception and Interaction Unit,
Athena Research and Innovation Center (Athena RIC)

# Part 4
# Multimodal Saliency and Video Summarization

**Athanasia Zlatintsi and Petros Koutras**

slides: http://cvsp.cs.ntua.gr/interspeech2018


Tutorial at INTERSPEECH 2018, Hyderabad, India, 2 Sep. 2018

# Outline

- Video Summarization, Human Attention & Saliency
- State-of-the-Art: Audio and Visual Saliency Approaches
- Multimodal Salient Event Detection: Methodology
- Multimodal Salient Event Detection: Results
  - COGNIMUSE database
- Applications & Demos

# Video Summarization
# Human Attention & Saliency

# Video Summarization

- Need for summarization
  - 400 hours of video are uploaded to YouTube every minute
  - Need to search for relevant content quickly in large amounts of video
- Summarization task refers to producing a shorter version of a video:
  - containing only the necessary and non redundant information required for context understanding
  - covering the interesting and informative frames or segments
  - without sacrificing much of the original enjoyability

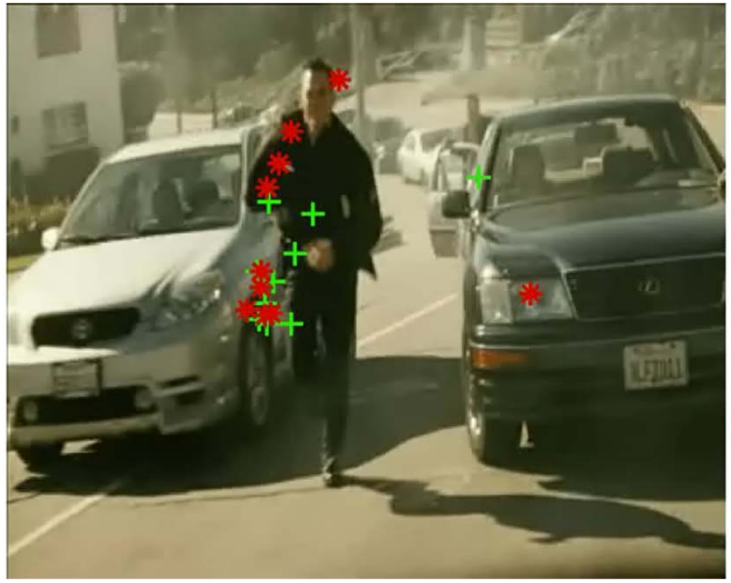
# Video Summarization Approaches

- Automatic summaries can be created with:
  - key-frames, which correspond to the most important video frames and represent a static storyboard

  - video skims that include the most descriptive and informative video segments

# Human Attention and Summarization

- Attention
  - Top-down, Task-driven
  - High level topics
- Saliency
  - Bottom-up, Data-Driven
  - Low level sensory cues



- Applications
  - Systems for selecting the most important regions/segments of a large amount of visual data
  - Video/Movie Summarization
  - Frontend for other applications like action recognition.

# Auditory Information Processing

- Our brain is capable of parsing information in the environment by using various cognitive processes
  - regardless various prominent distractors (known as the ‘cocktail party problem’)
- Such processes allow us to navigate to the soundscape, focus on conversations of interest, enjoy the background music and be alert to any **salient** sound events, i.e., when someone is calling us

[T. Darrell, J. W. Fisher III, P. Viola and W. Freeman, *Audio-visual Segmentation and “The Cocktail Party Effect”*, ICMI 2000]
[C. Alain and L. Bernstein, *From sounds to meaning: the role of attention during auditory scene analysis*, Curr. Opin. Otolaryngol. Head Neck Surg. 16, 2008]

# Auditory Attention

Adjusted by:

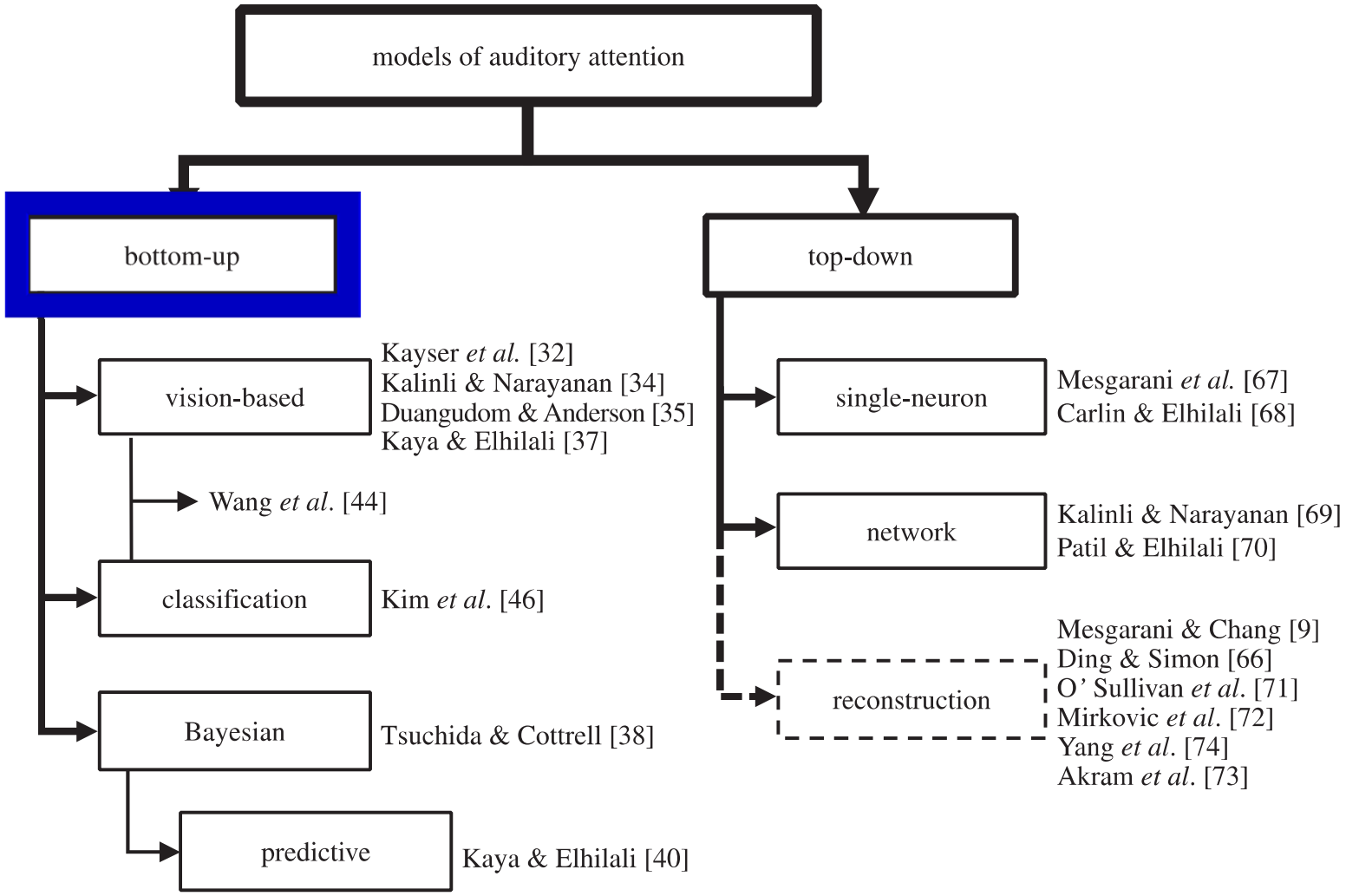
- 'Bottom-up' sensory-driven task-independent factors (automatic, 'unconscious', stimulus driven, little or no attention)
- 'Top-down' task-dependent goals, expectations and learned schemas ('conscious', effortful, selective, memory dependent)
- It acts as a selection process that focus both sensory and cognitive resources on the most relevant events in the soundscape, i.e.,
  - a sudden loud explosion
  - or a task at hand, e.g., listen to announcements in a busy airport

# Auditory Saliency

- Quality to stand out relatively to the surrounding soundscape
  - Salient stimuli are able to attract our attention and are easier to detect
- Describes the potential influence of a stimulus on our perception and behavior
- Key attentional mechanism facilitating learning and survival
- Complements the frequently studied processes of attention and detection
- Introduces a qualitative description of those stimulus properties relevant for these processes

# State-of-the-Art Audio and Visual Saliency Approaches

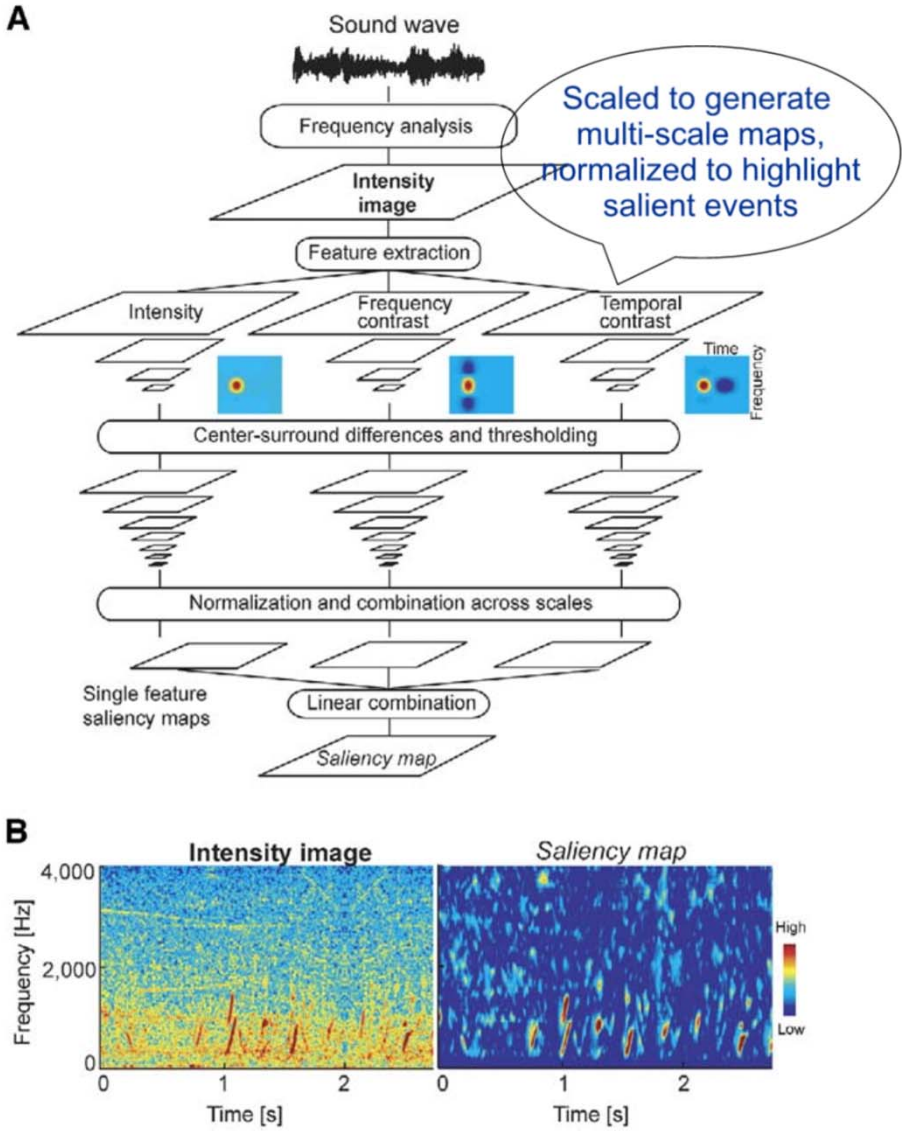
# Auditory Saliency and Attention: Approaches




[E.M. Kaya and M. Elhilali, *Modelling auditory attention*. Phil. Trans. R. Soc., 2016]

# An Auditory Saliency Map

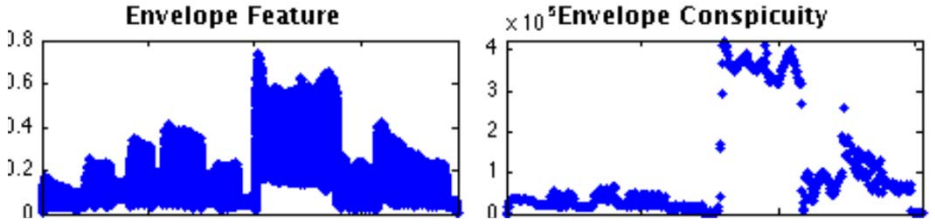
- Time - frequency representation of the sound waveform as an “auditory image”
- Spectro-temporal features such as intensity, frequency & temporal contrast
- The model was able to match both human and monkey behavioral responses for the detection of salient sounds in noisy scenes
- Demonstrated that saliency processing in the brain may share commonalities across sensory modalities
- Provided a guide for the design of psychoacoustical experiments to probe auditory bottom-up attention in humans



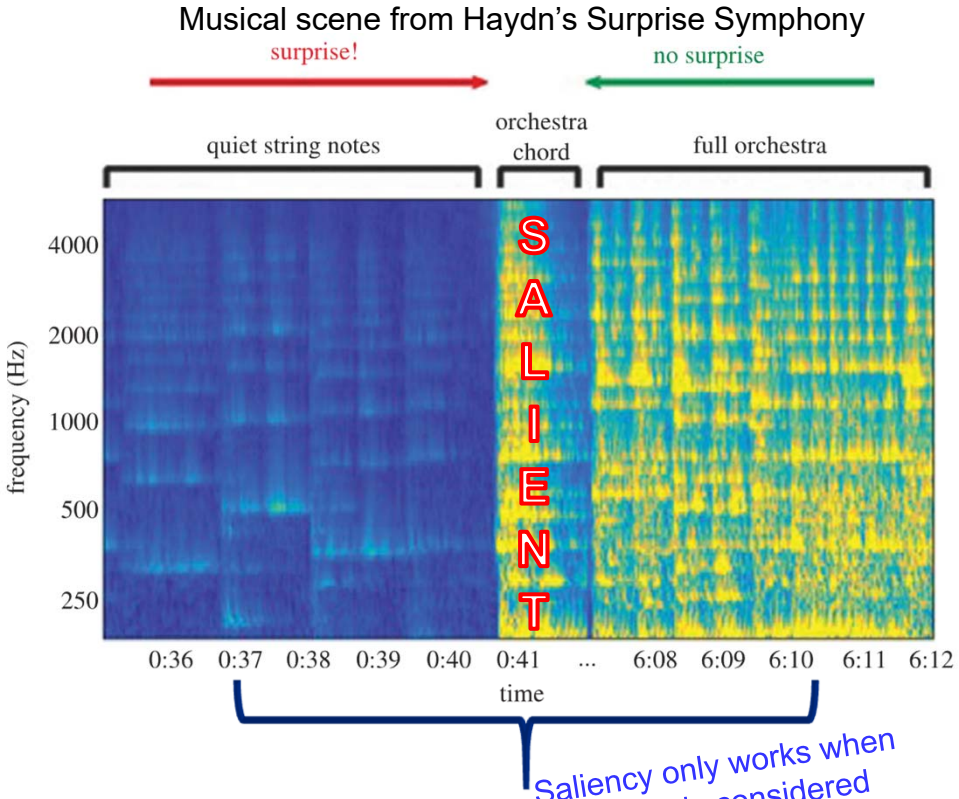

[C. Kayser, C.I. Petkov, M. Lippert and N.K. Logothetis, *Mechanisms for allocating auditory attention: an auditory saliency map*, Curr. Biol., 2005]

# Temporal Saliency Map

- Based on the fact that sound is actually a naturally temporally evolving entity
- Saliency is measured by how much an event differs from its surrounding, thus. sounds preceding in time
- Auditory scene is treated as single dimensional temporal input (at all times), rather than as an image
- Employing perceptual properties of sound, i.e., loudness, pitch and timbre
- Feature analysis over time to highlight their dynamic quality before normalizing and integrating across feature maps

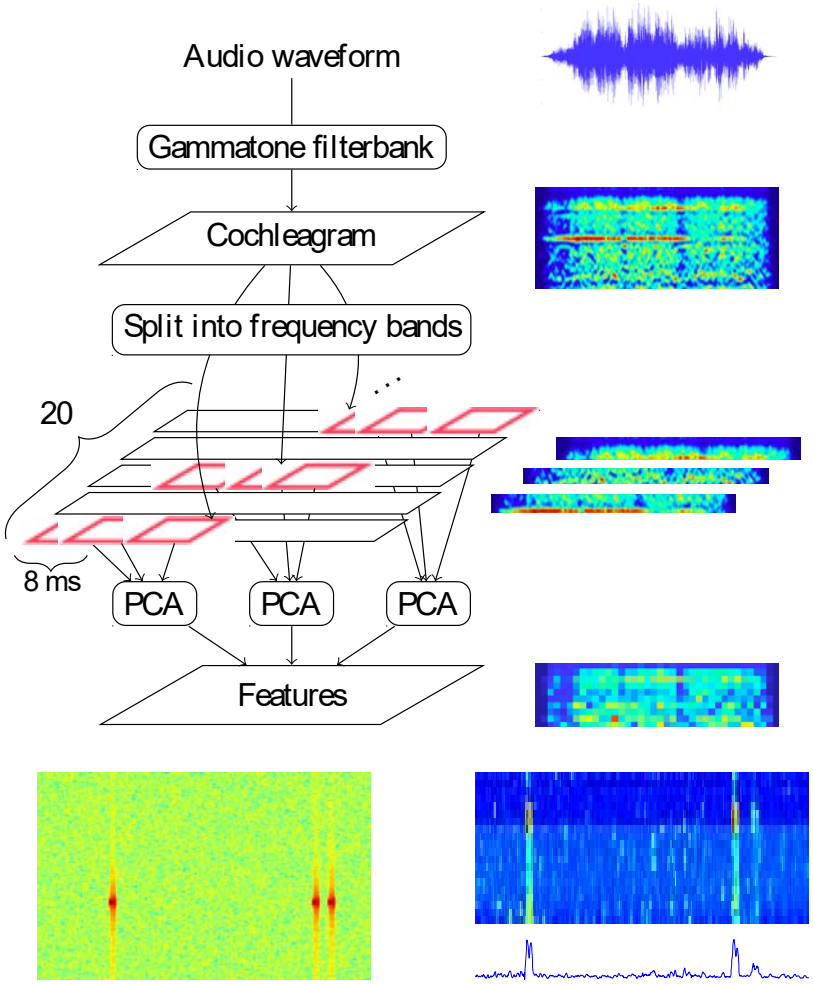



Differences in saliency selection between the temporal saliency model and Kayser's saliency model




[E.M. Kaya and M. Elhilali, *A temporal saliency map for modeling auditory attention*, CISS 2012]
[E.M. Kaya and M. Elhilali, *Modelling auditory attention*. Phil. Trans. R. Soc., 2016]

# Statistical-based Approach



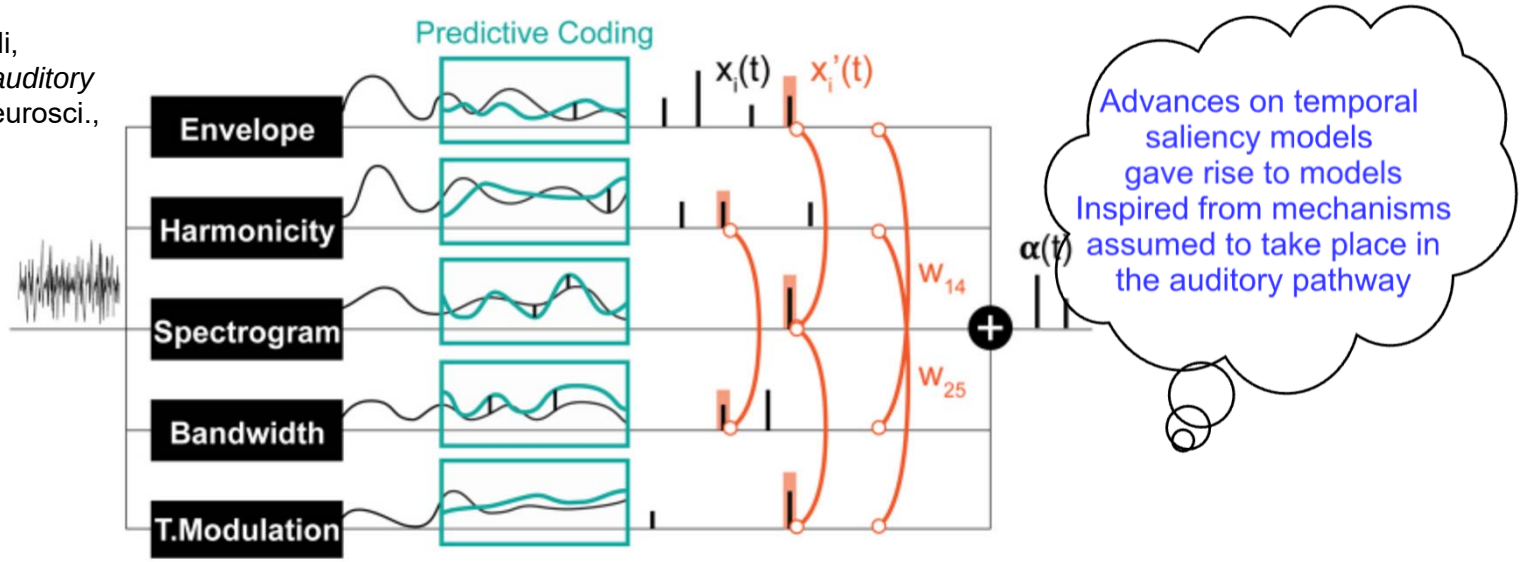

Spectrogram and saliency map (saliency values summed over frequency axis) for single and paired tones.

- Statistical approach adapted from vision*
- Combination of long-term statistics computed using natural sounds with short-term, temporally local, rapidly changing statistics of the incoming sound
- A sound is flagged as **salient** if it is determined to be unusual relative to learned statistics
- Cochleogram was used instead of a spectrogram (for computational efficiency) and PCA for dimensionality reduction


[T. Tsuchida and G. Cottrell, *Auditory saliency using natural statistics*, Society for Neuroscience Meeting, 2012]
[*L. Zhang, M.H. Tong, T.K. Marks, H. Shan and G.W. Cottrell, *SUN: A Bayesian framework for saliency using natural statistics,* J. Vis., 2008]

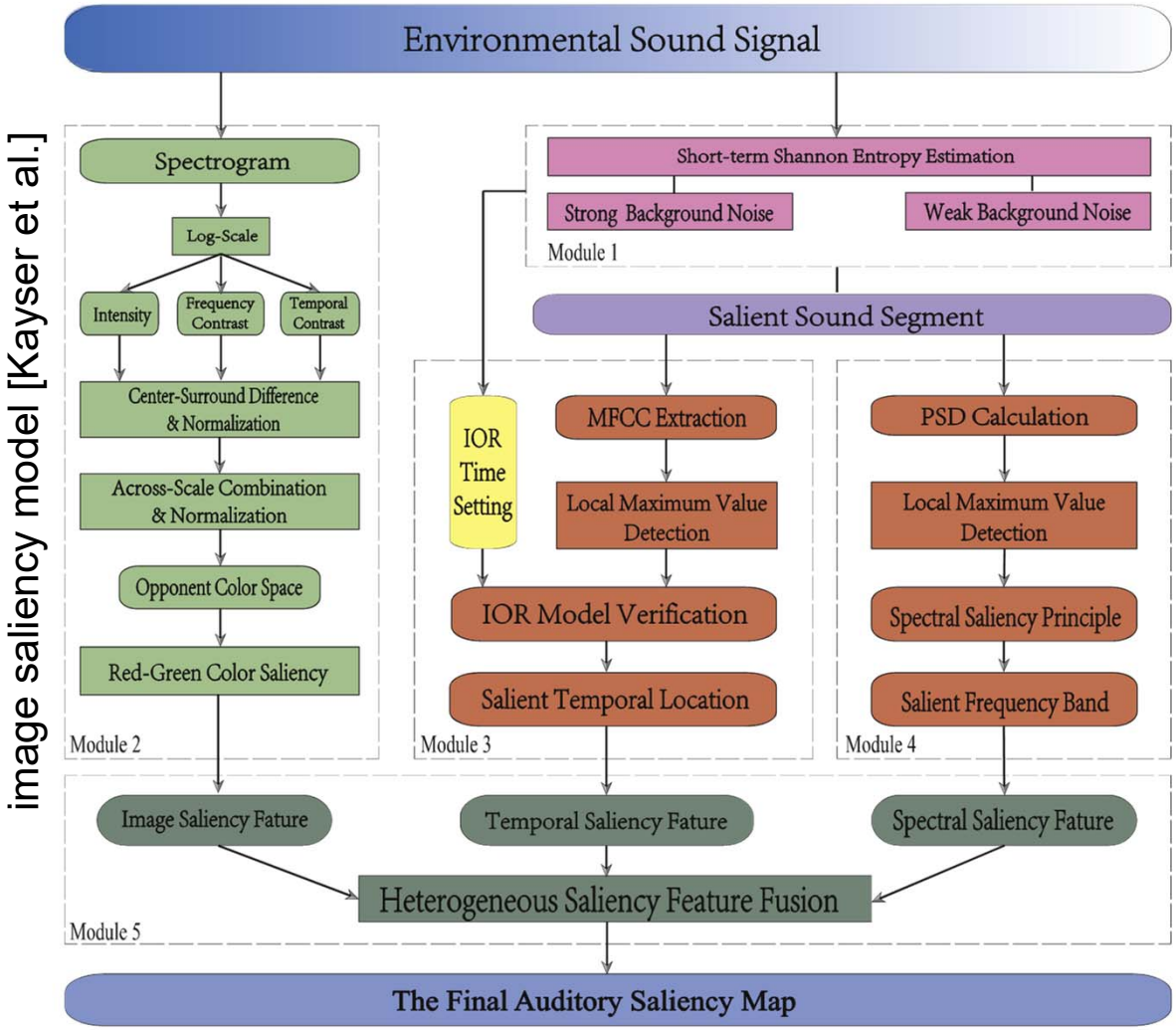
# Predictive Coding

[E.M. Kaya and M. Elhilali, *Investigating bottom-up auditory attention*, Front. Hum. Neurosci., 2014]




- Emphasis on the role of processing events over time and shaping neural responses of current sounds based on their preceding context
- Mapping of the acoustic waveform onto a high-dimensional auditory space, encoding perceptual loudness, pitch and timbre of the incoming sound, building upon evolving temporal features
- Collect feature statistics over time and make predictions about future sensory inputs
- Regularities are tracked, and deviations from regularities are flagged as **salient**
- Nonlinear interaction across features, using asymmetrical weights between pairwise features

# Bio-inspired Saliency Detection




**Composite system**:

- Shannon Entropy for global saliency: measure the sound's informational value
- MFCCs for acoustic saliency: temporal analysis of sound characteristics (using IOR model for saliency verification)
- Spectral saliency: analysis of the power spectral density of the stimulus
- Kayser's image model
- Robustness to saliency estimation especially in noisy environements

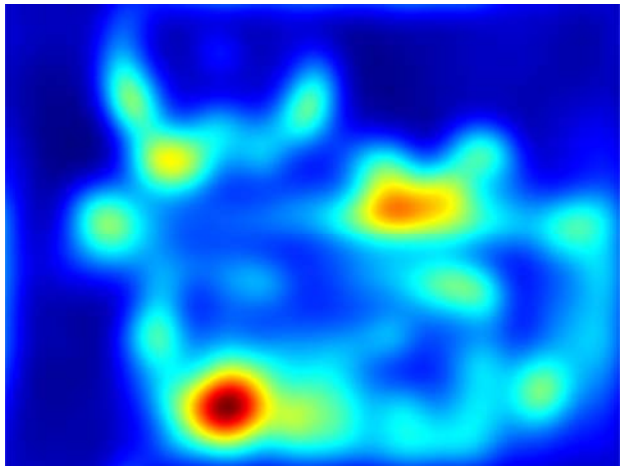
[J. Wang, K. Zhang, K. Madani and C. Sabourin, *Salient environmental sound detection framework for machine awareness*, Neurocomp. 2015]

# Visual Saliency: Approaches, Evaluation

- Spatial Saliency
  - predict viewers fixations in image plain
  - static eye-tracking datasets: Toronto data set, MIT CAT200, SALICON, …
- Spatio-Temporal Saliency
  - predict viewers fixations both in space and time
  - dynamic eye-tracking datasets: CRCNS, DIEM, DHF1K, …
- Temporal Saliency
  - find the frames or segments that contain the most salient events
  - visual, audio and text streams → multimodal salient events
  - human annotated databases: COGNIMUSE multimodal video database

Original Image

Spatial Saliency Map



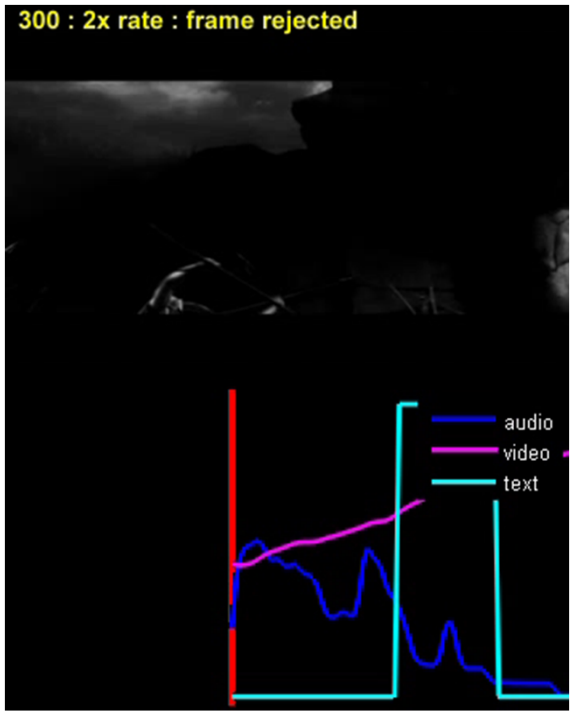
# Visual Saliency: Approaches, Evaluation

- Spatial Saliency
  - predict viewers fixations both in image plain
  - static eye-tracking datasets: Toronto data set, MIT CAT200, SALICON, …
- Spatio-Temporal Saliency
  - predict viewers fixations both in space and time
  - dynamic eye-tracking datasets: CRCNS, DIEM, DHF1K, …
- Temporal Saliency
  - find the frames or segments that contain the most salient events
  - visual, audio and text streams → multimodal salient events
  - human annotated databases: COGNIMUSE multimodal video database

Spatio-Temporal Saliency Map

# Visual Saliency: Approaches, Evaluation

- Spatial Saliency
  - predict viewers fixations both in image plain
  - static eye-tracking datasets: Toronto data set, MIT CAT200, SALICON, …
- Spatio-Temporal Saliency
  - predict viewers fixations both in space and time
  - dynamic eye-tracking datasets: CRCNS, DIEM, DHF1K, …
- Temporal Saliency
  - find the frames or segments that contain the most salient events
  - visual, audio and text streams → multimodal salient events
  - human annotated databases: COGNIMUSE multimodal video database

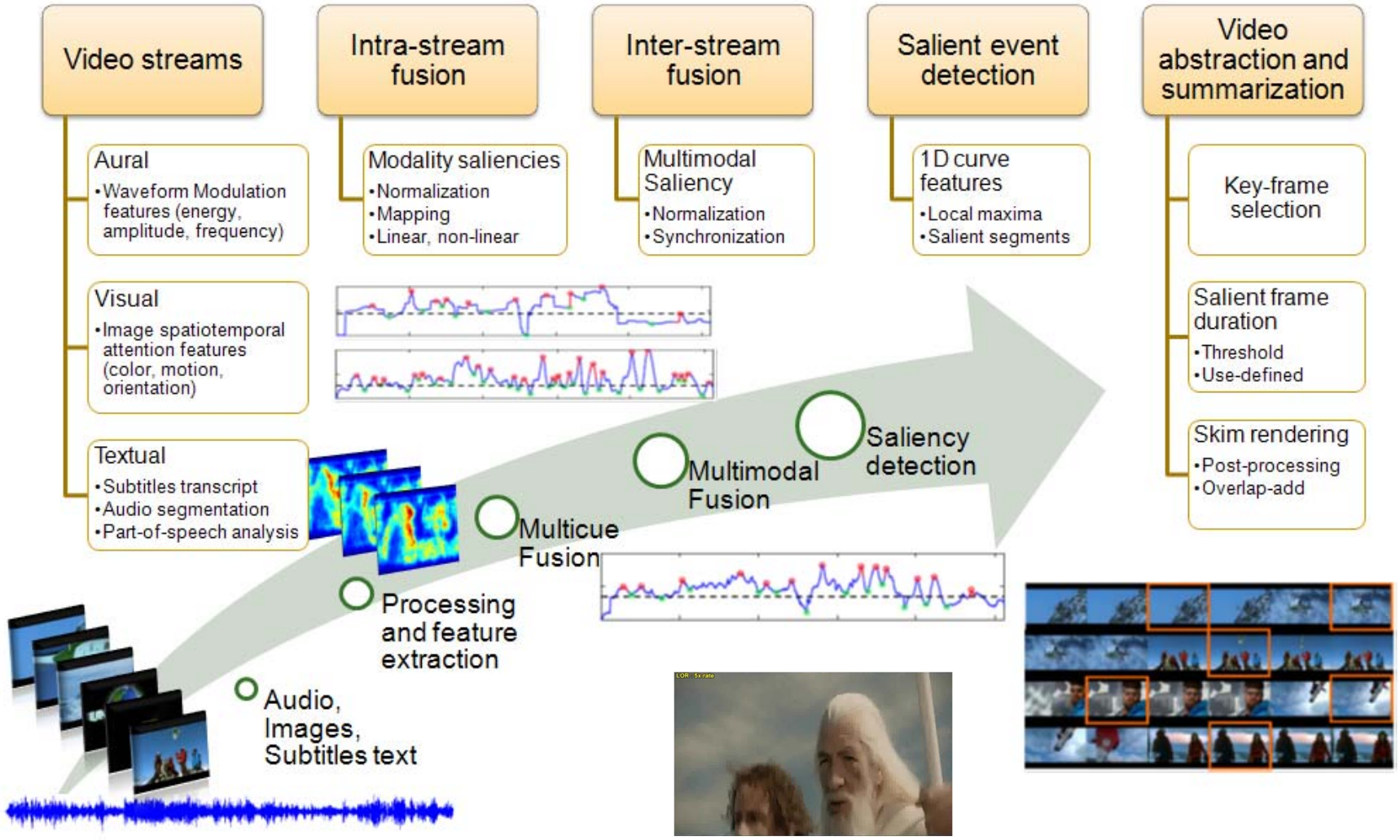


# Multimodal Salient Event Detection: Methodology

# Multimodal Saliency & Movie Summarization

**COGNIMUSE: Multimodal Signal and Event Processing In Perception and Cognition**

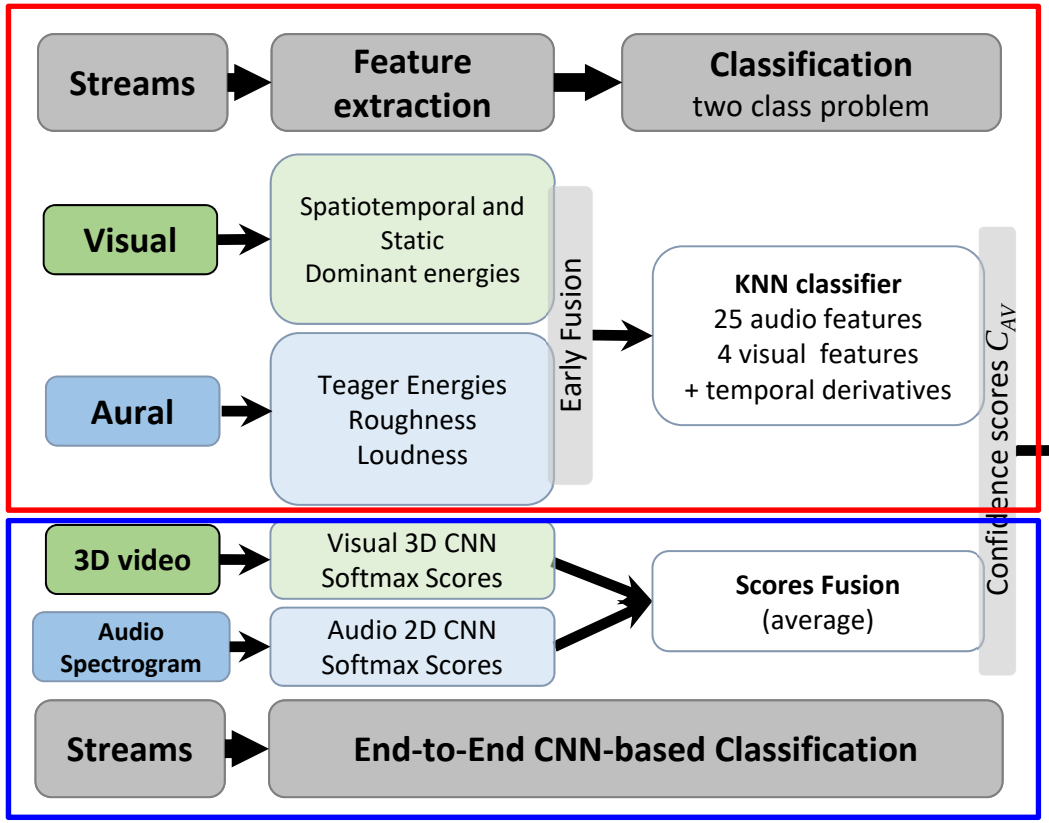
website: http://cognimuse.cs.ntua.gr/



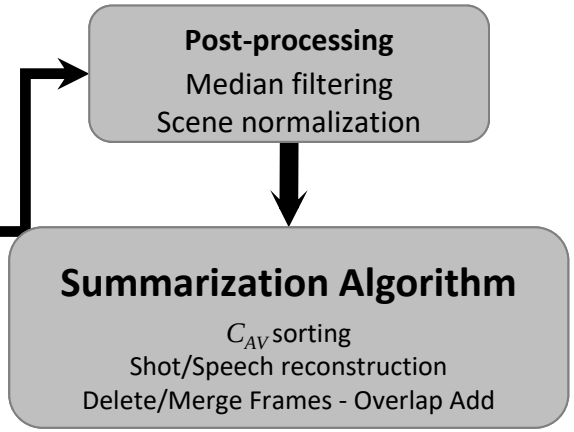

[G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas and Y. Avrithis, *Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention,* IEEE Transactions on Multimedia, vol. 15, no. 7, Nov. 2013.]

# Multimodal Salient Event Detection:
## Handcrafted vs. Multimodal CNN-based approach

handcrafted features + classification algorithms




multimodal CNN-based architectures for saliency detection

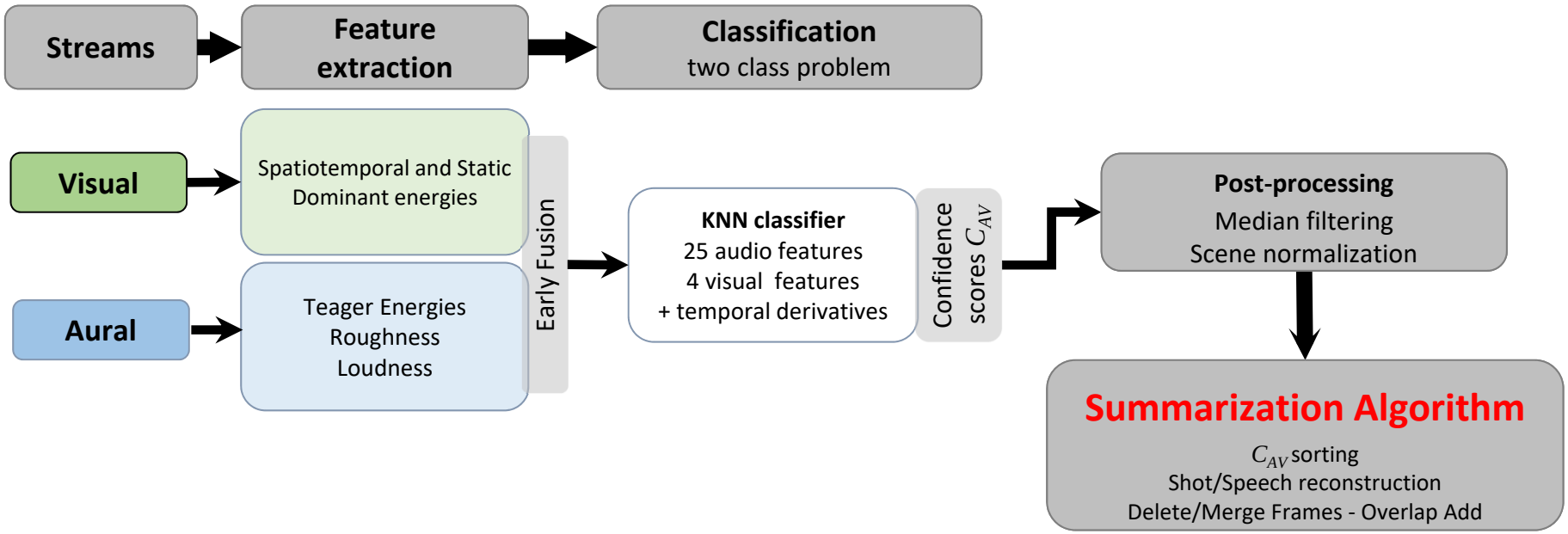



experimental evaluation

COGNIMUSE database:
- multimodal video database with multilevel saliency annotation
- *http://cognimuse.cs.ntua.gr/database*

[P. Koutras, A. Zlatintsi and P. Maragos, *Exploring CNN-based architectures for Multimodal Salient Event Detection in Videos*, IVMSP 2018.]

# Handcrafted Frontend




- Signal processing methods for feature extraction
  - unified energy-based framework for audio-visual saliency
  - perceptually inspired and carefully designed filterbanks
- Machine learning algorithms for salient events classification
- Postprocessing of the final scores

[P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos and A. Potamianos, *Predicting Audio-visual Salient Events based on A-V-T Modalities For Movie Summarization*, ICIP 2015.]

# Handcrafted Audio Analysis Overview

Teager-Kaiser Operator: $\Psi[t] = \dot{x}^2 - x\ddot{x}$

AM-FM Modulated Audio Signal *(narrow-band)*:

$$x(t) = \alpha(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)$$

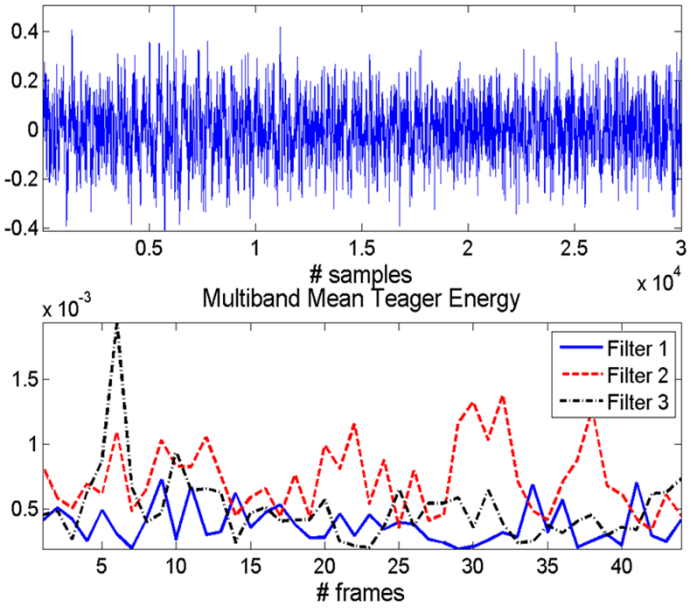
$$\frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \simeq |\alpha(t)| \qquad \frac{\sqrt{\Psi[\dot{x}(t)]}}{\sqrt{\Psi[x(t)]}} \simeq \omega(t)$$

narrow-band→ Filterbank of 25 Mel arranged Gabor filters

## AM-FM model

## Auditory Features

- 25 Gabor Energies
- Roughness
- Loudness




- **Roughness** (or sensory dissonance)
  - expresses the "stridency" of a sound due to rapid fluctuations in the amplitude
- **Loudness** (perceived sound pressure level)
  - Loudness model for time-varying sounds by Zwicker & Fastl (1999)

[A. Zlatintsi, E.Iosif, P. Maragos and A. Potamianos. *Audio Salient Event Detection and Summarization using Audio and Text Modalities*, EUSIPCO, 2015]

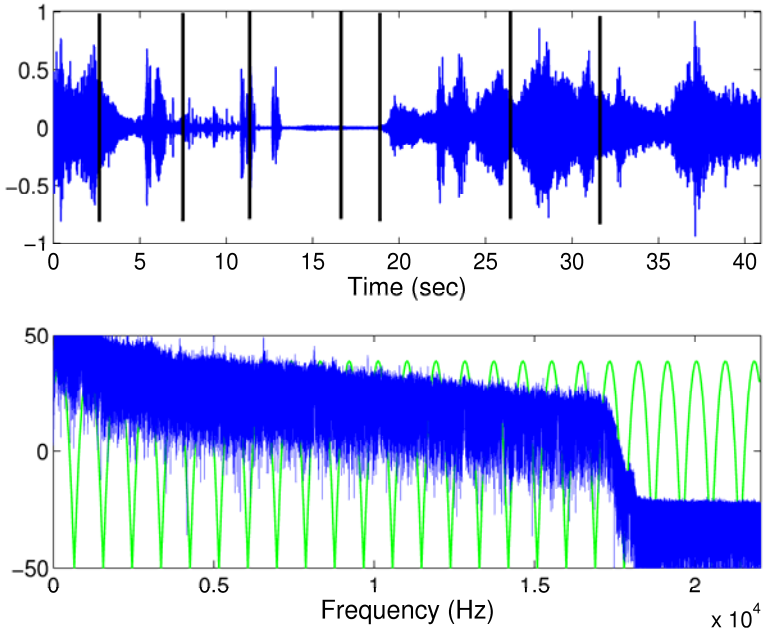
# AM-FM Model

- AM-FM model for audio signals

  (speech, music environmental sounds)

$$s(t) = \sum_{i=1}^{K} \alpha_i(t) \cos(\varphi_i(t))$$

instantaneous amplitude — phase




Modeling of each resonance component as an amplitude and frequency modulated sinusoid (AM-FM signal), and the whole signal as a sum of such AM-FM components

- Teager-Kaiser energy operator for nonlinear energy tracking
- ESA (Energy Separation Algorithm) for AM-FM demodulation

- Multiband filtering with:
  - *K* Gabor filters $h_k$, narrowband components

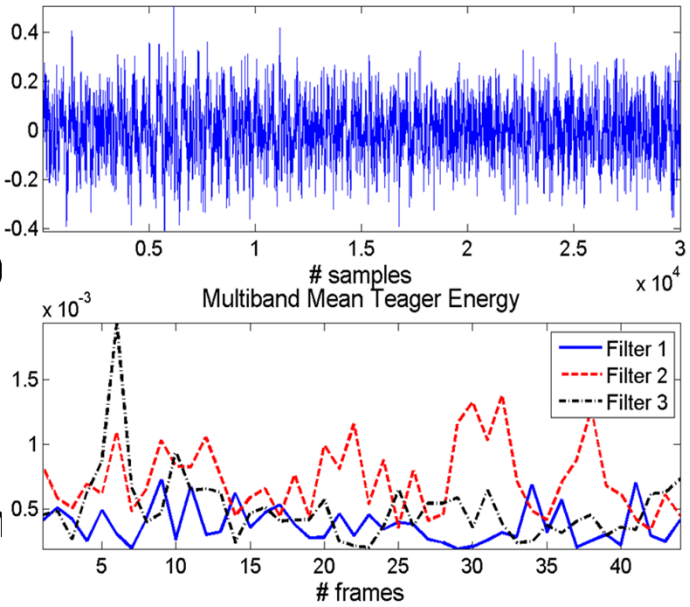
[P. Maragos, J.F. Kaiser and T.F. Quatieri, *Energy Separation in Signal Modulations with Application to Speech Analysis,* IEEE Trans. on Signal Process., 1993]

# Teager-Kaiser Operator

## Teager-Kaiser Energy Operator:

$$\Psi[x] = \dot{x}^2 - x\ddot{x}, \text{ where } \dot{x} = dx/dt$$

- For energy estimation and AM-FM demodulation, using **ESA** (Energy Separation Algorithm)
- Captures amplitude and frequency variation information
- Ψ can detect robustly & discriminate various acoustic events due to its sharp time resolution and lowpass behavior
- Important for auditory scene analysis
- Robust to noise compared to the squared energy operator
- Multiband TECC (Teager Energy Cepstrum Coefficients) successful in speech recognition

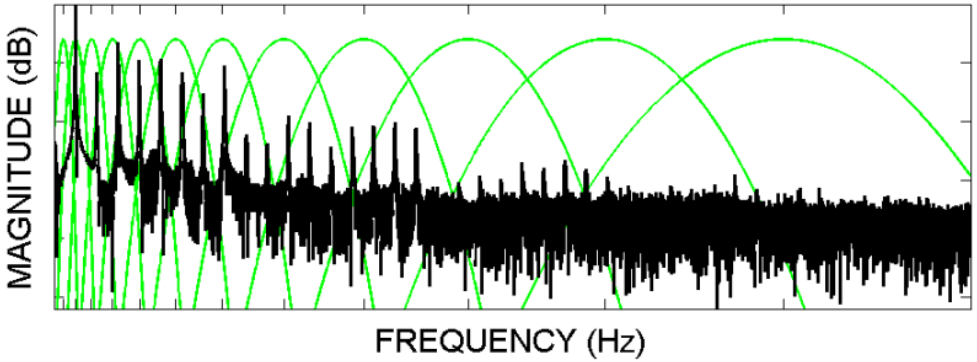



[J.F. Kaiser, *On a simple algorithm to calculate the energy of a signal*, ICASSP 1990]
[D. Dimitriadis, P. Maragos and A. Potamianos, *On the effects of filterbank design and energy computation on robust speech recognition*, TASLP 2011]

# Multiband Filtering

**Teager Energy**: only meaningful in narrowband signals

- **Multiband filtering** of the signal with **Gabor filters**
  - Gabor filtering for isolation of narrowband components
  - Gabor filters: exhibit good joint time-frequency resolution

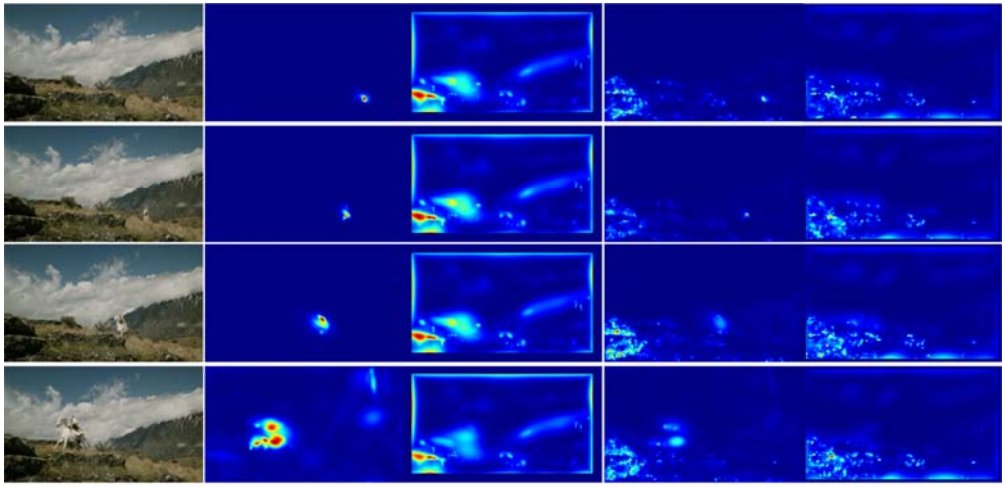


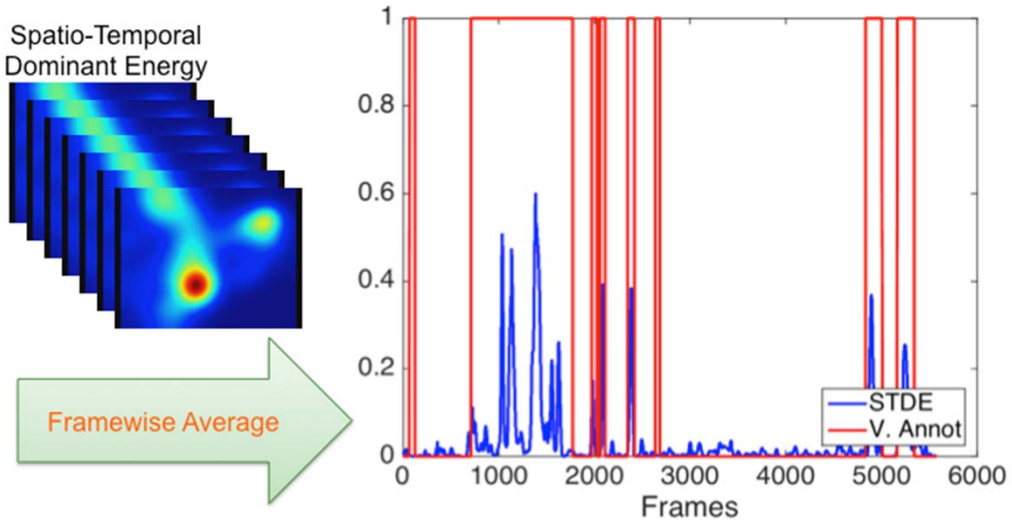
# Handcrafted Visual Saliency Model

## 3D Gabor Energy model

### Visual Features

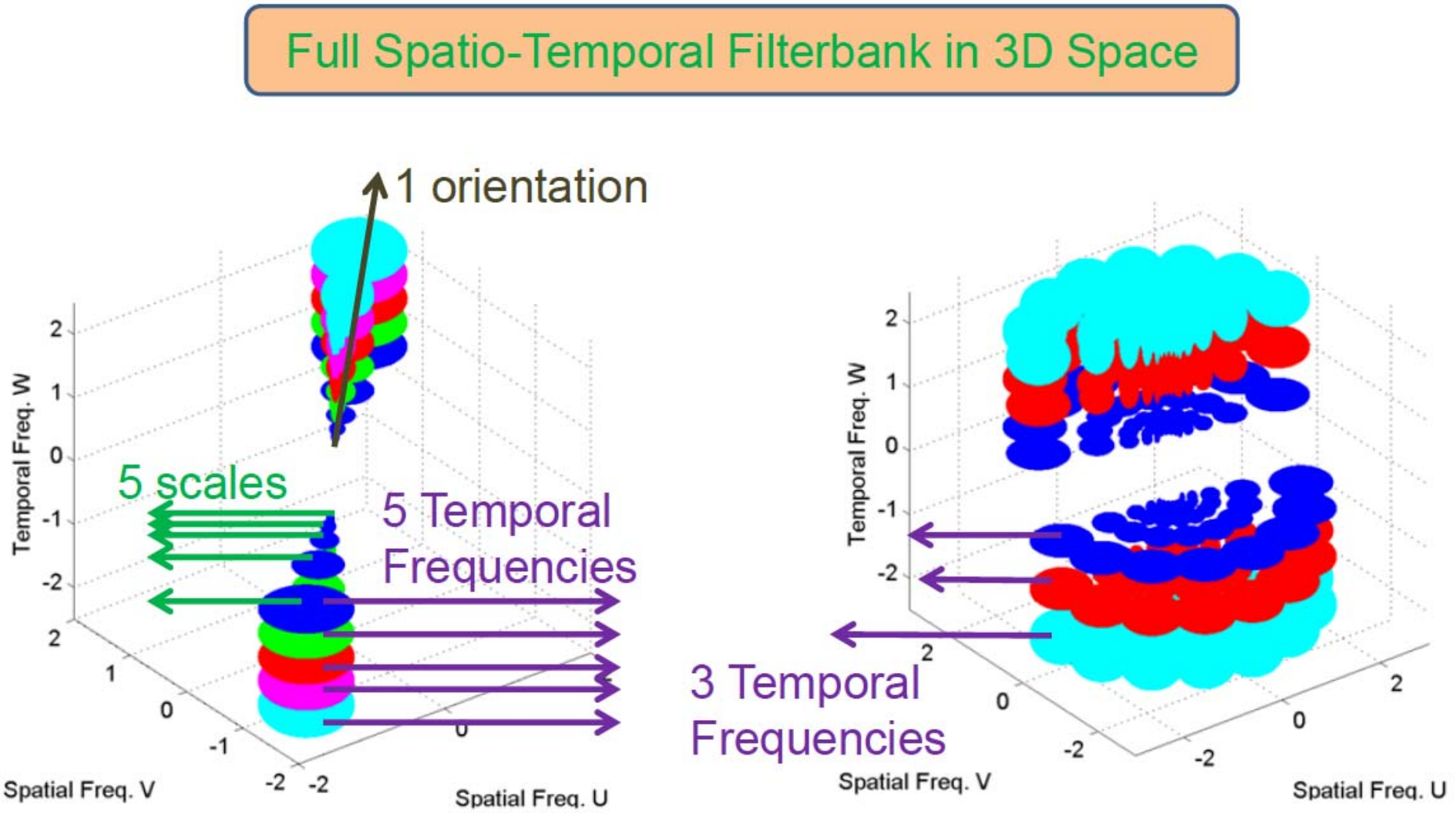
- Both luminance and color streams:
  - Spatio-Temporal Dominant Energies (Filterbank of 400 3D Gabor filters)
  - Spatial Dominant Energies (Filterbank of 40 Spatial Gabor filters)
- Energy Curves
  - Simple 3D to 1D Mapping
  - Mean value for each 2D frame slice of each 3D energy volume
  - 4 temporal sequences of visual feature vectors.



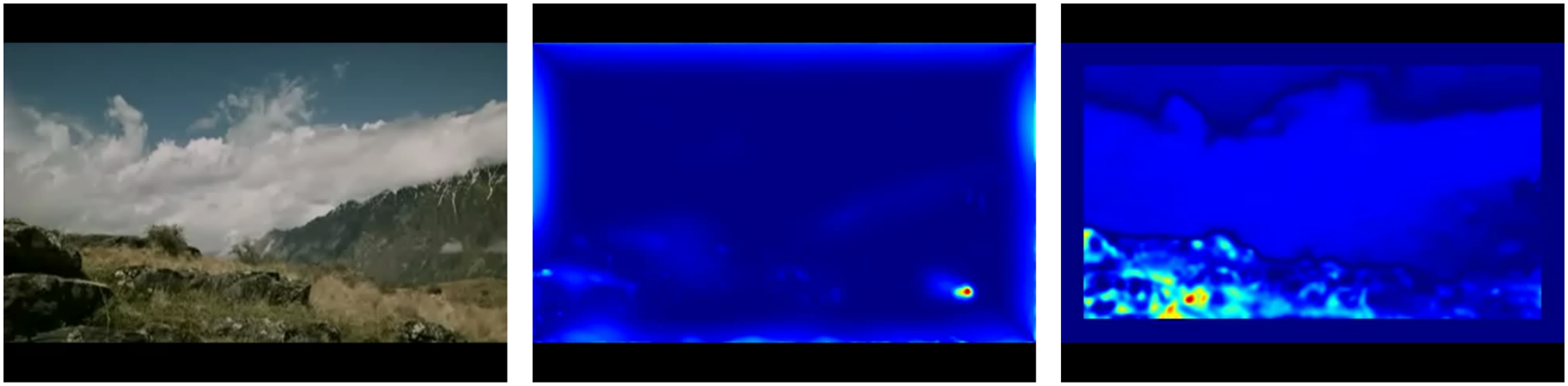

[P. Koutras and P. Maragos. *A Perceptually-based Spatio-Temporal Computational Framework for Visual Saliency Estimation*, Signal Proc.: Image Comm, 2015]

# Spatio-Temporal Gabor Filterbank



# Visual Saliency in Movie Videos - Demo

Original RGB Frames | Luminance STDE | Color SDE

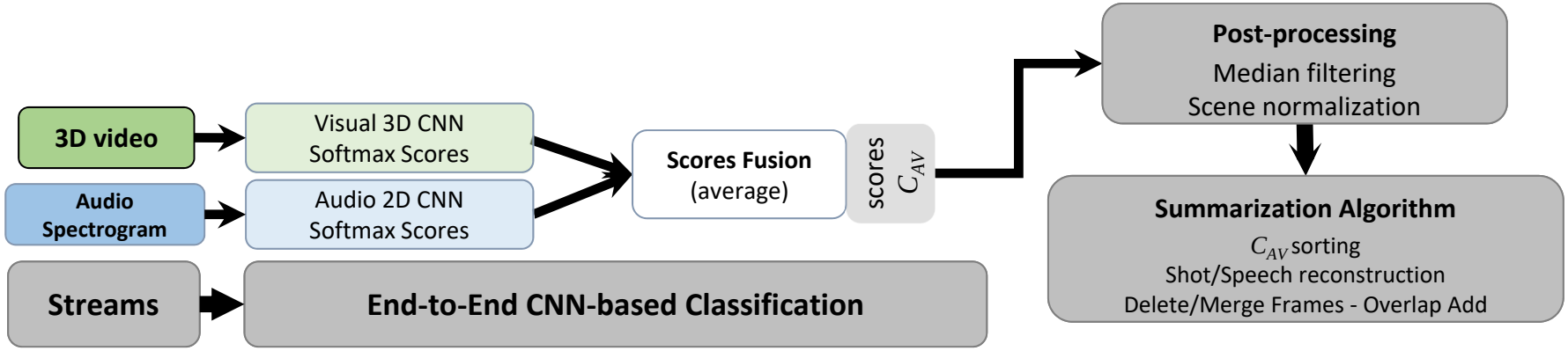


COGNIMUSE Database:
Lord of the Rings: The Return of the King

# Salient Events Classification

- Multimodal features vectors:
  - standardize features (zero mean, unit covariance)
  - compute 1st and 2nd order derivatives (deltas)
  - concatenate audio and visual features
- Classification based conventional machine learning:
  - binary classification problem (salient / non salient video segments)
  - K-Nearest Neighbor Classifier (KNN)
    - confidence scores for every classification result
    - continuous indicator function curve → represents the most salient events
- Scores postprocessing
  - median filtering and scene normalization of saliency measurement
  - sorting the frames based on saliency measurement
  - summarization algorithm

# CNN-based Architectures for Saliency Detection

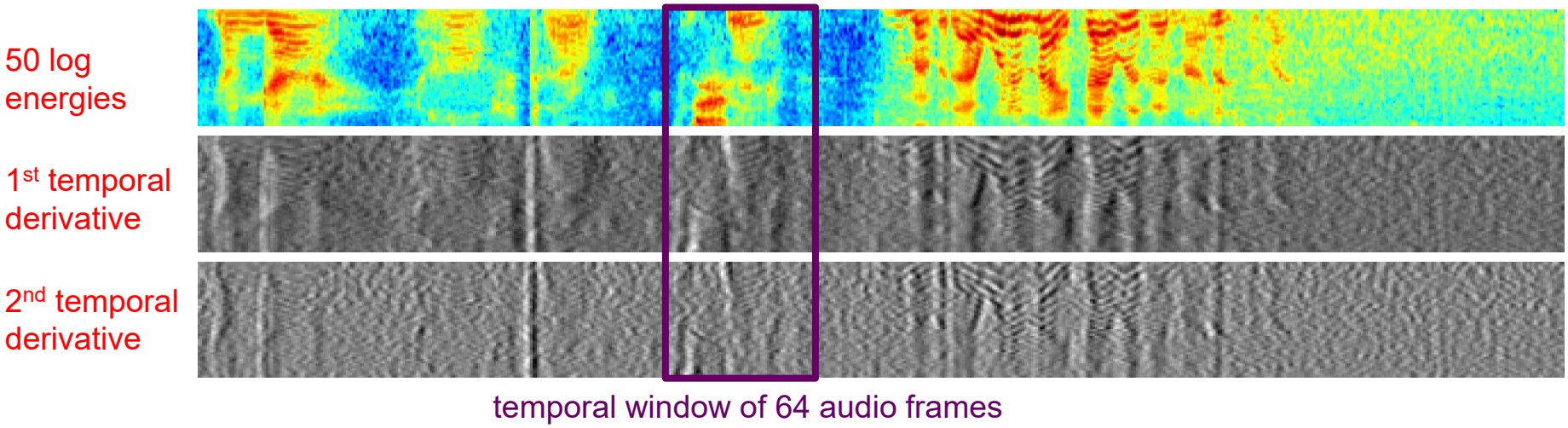



- Two-steam Convolutional Networks: video and audio
  - inspired from two-stream CNNs for action recognition (RGB + Flow, RGB + Depth)
  - replace the stages of feature extraction and classification with one single network
  - softmax scores of the CNN output as visual and audio saliencies
- Use monomodal or multimodal annotations as ground-truth labels
  - multinomial logistic loss for binary classification:

$$\mathcal{L}(\mathbf{W}) = -\sum_{j \in Y_+} \log P(y_j = 1 | X; \mathbf{W}) - \sum_{j \in Y_-} \log P(y_j = 0 | X; \mathbf{W})$$

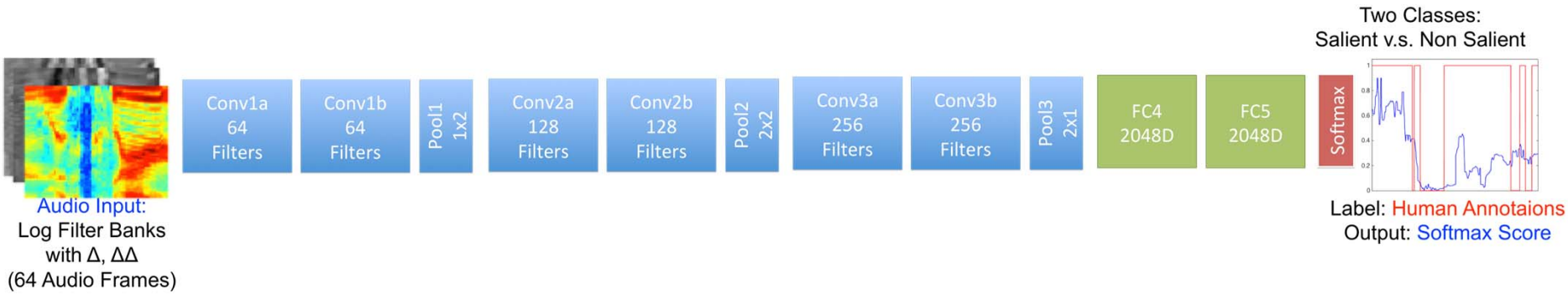
- Employ trained models for computing saliency curves in a new video
  - same postprocessing of the final scores as in handcrafted frontend

# Audio 2D Time-Frequency Representation




- Represent the raw audio signal in the 2D time-frequency domain
  - preserve locality in both time and frequency axis
  - conventional MFCCs representation cannot maintain locality to frequency axis due to the DCT projection
- Employ log-energies using 25 ms frames with 10 ms shift
  - compute first and second temporal derivatives
- Temporal segments of 64 audio frames
  - synchronized with the 16-frames video clips

# CNN Architecture for Auditory Saliency

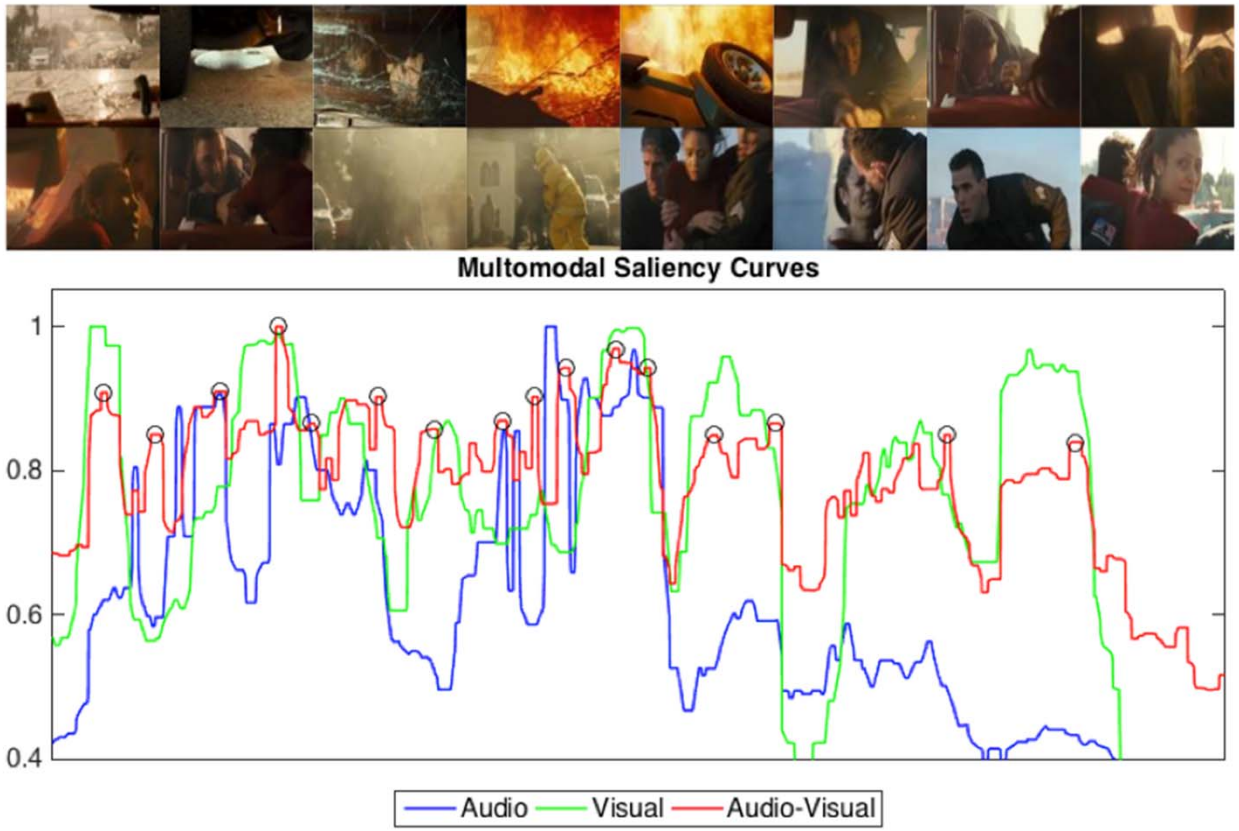



- Deep 2D CNN architecture
  - input: 3 channel 2D input, similarly to the RGB image
    - synchronized with visual clips
  - output: softmax score as auditory saliency curve
- 2D convolutional and 2D max-pooling operations over time and frequency
  - based on the VGG idea of small kernels
- Train end-to-end using the audio-only or the audio-visual human annotation

# CNN Architecture for Visual Saliency


- Deep end-to-end CNN architecture
  - input: split video into 16-frame RGB clips
    - total ~18000 clips for training
  - output: softmax score as visual saliency curve
- Learn filterbank parameters as a sequence of 3D convolutional networks (C3D)
  - convolutions and pooling operations are applied inside spatio-temporal cuboids
- Learn spatio-temporal patterns related to visual saliency
- Train end-to-end using the visual-only or the audio-visual human annotation

# CNN Estimated Audio-Visual Saliency Curves




- Audio-Visual Saliency Curves
  - two-stream CNNs trained with the audio-visual annotation labels
  - average the softmax scores
- Keyframes extracted as local extrema of the audio-visual curve

# COGNIMUSE Database
## Saliency, Semantic & Cross-Media Events Database

*http://cognimuse.cs.ntua.gr/database*

Including:

- Saliency annotation on multiple layers
- Audio & Visual events annotation
- COSMOROE cross-media relations annotation
- Emotion annotation

Database Content:

- **7** 30-min **movie clips** from: Beautiful Mind (BMI), Chicago (CHI), Crash (CRA), The Departed (DEP), Gladiator (GLA), Lord of the Rings III: The return of the king(LOR), Finding Nemo (FNE)
- **5** 20-min **travel documentaries**
- **1** 100-min **movie**: Gone with the Wind (GWTW)

[A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Marandrakis, N. Efhymiou, K. Pastra, A. Potamianos and P. Maragos, *COGNIMUSE: A Multimodal Video Database Annotated with Saliency, Events, Semantics and Emotion with Application to Summarization,* EURASIP Jour. on Image and Video Proc., 2017]
[A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos and K. Pastra, *Quality Evaluation of Computational Models for Movie Summarization,* QoMEX 2015]

# Database Annotation: Saliency & Structure

**Movie Structure:**

- Shots 370-699 (~540/movie)
- Scenes 7-23 (~14/movie)

**Generic Sensory Saliency:**

1) Audio-only
2) Visual-only
3) Audio-Visual (AV)

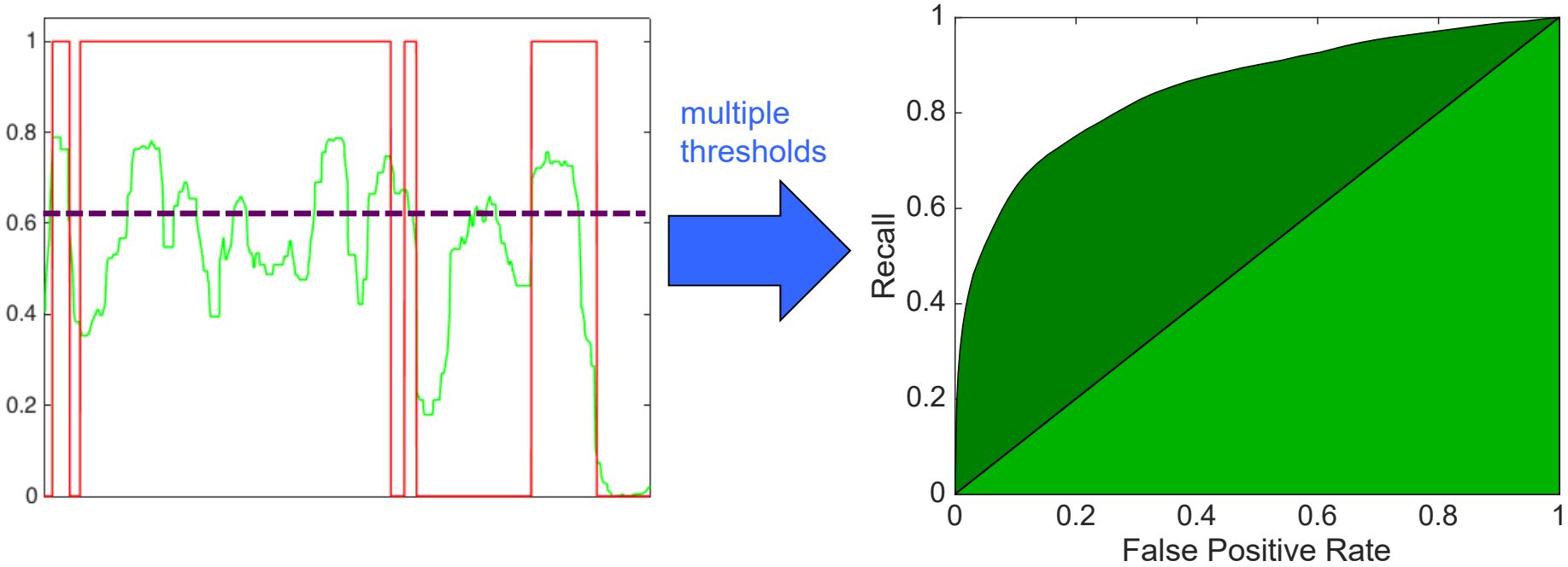
- Based on movie elements that capture the viewers' attention instantaneously or in segments
- Done quickly/effortlessly & without any focused attention or though
- Little or no searching required

**Attentive Saliency (Cognitive Attention):**

1) Semantics: segments that are conceptually important, e.g., phrases, actions, symbolic information, sounds….

# Multimodal Salient Event Detection: Results

# Salient Event Detection: Evaluation Metric




- Compare continuous saliency curves with binary annotations
- Area Under Curve (AUC)
  - apply threshold to saliency curves
  - area under the Receiver Operating Characteristic (ROC) curve (False Positive Rate – Recall)
  - binary classification problem: (salient / non salient segments)

# Evaluation Results – Hollywood Movies

A Beautiful Mind

Gladiator

Chicago

Lord of the Rings

Crash

The Departed

| AUC Results | V-V | | A-A | | AV-AV (mean) | |
|---|---|---|---|---|---|---|
| videos | Hndcr. | CNN | Hndcr. | CNN | Hndcr. | CNN |
| Six Hollywood Movies | | | | | | |
| BMI | 0.718 | 0.765 | 0.823 | 0.844 | 0.842 | 0.839 |
| GLA | 0.739 | 0.772 | 0.840 | 0.849 | 0.850 | 0.830 |
| CHI | 0.645 | 0.706 | 0.847 | 0.815 | 0.819 | 0.820 |
| LOR | 0.688 | 0.738 | 0.873 | 0.872 | 0.811 | 0.832 |
| CRA | 0.720 | 0.726 | 0.848 | 0.874 | 0.804 | 0.799 |
| DEP | 0.778 | 0.741 | 0.822 | 0.861 | 0.824 | 0.856 |
| Aver. | 0.715 | **0.742** | 0.842 | **0.853** | 0.825 | **0.830** |

- Six fold cross-validation
  - five movies were used for training and tested on the sixth
- CNN-based architecture outperforms the hand-crafted frontend
  - for audio modality only in CHI we cannot achieve improvement:
    - musical containing mostly music segments
    - CNN training on the other movies that do not contain this information

# Evaluation Results – Travel Documentaries

AR - London

AR - Rio

AR - Sydney

AR - Tokyo

GoT - London

| AUC Results | V-V | | A-A | | AV-AV (mean) | |
|---|---|---|---|---|---|---|
| videos | Hndcr. | CNN | Hndcr. | CNN | Hndcr. | CNN |
| Five Travel Documentaries | | | | | | |
| LON | 0.650 | 0.806 | 0.794 | 0.830 | 0.777 | 0.814 |
| RIO | 0.668 | 0.718 | 0.690 | 0.737 | 0.821 | 0.805 |
| SYD | 0.621 | 0.771 | 0.726 | 0.787 | 0.734 | 0.863 |
| TOK | 0.767 | 0.831 | 0.796 | 0.849 | 0.819 | 0.856 |
| GLN | 0.657 | 0.679 | 0.809 | 0.894 | 0.693 | 0.810 |
| Aver. | 0.673 | **0.761** | 0.763 | **0.819** | 0.769 | **0.830** |

- Five fold cross-validation
  - four movies were used for training and the fifth for testing
- CNN-based architecture outperforms the hand-crafted frontend
  - Greater improvements for visual modality

# Evaluation Results – Full Length Movie

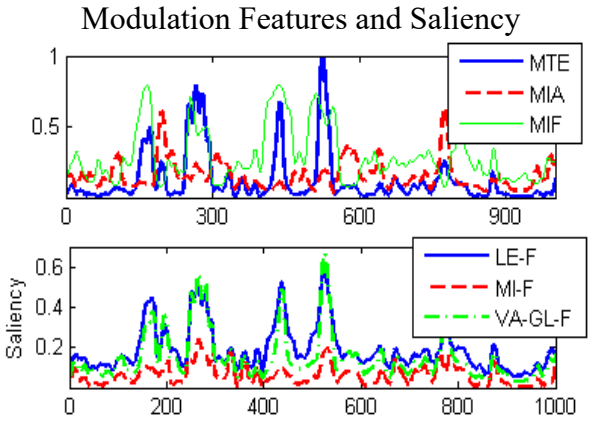
Gone with the Wind – Part 1

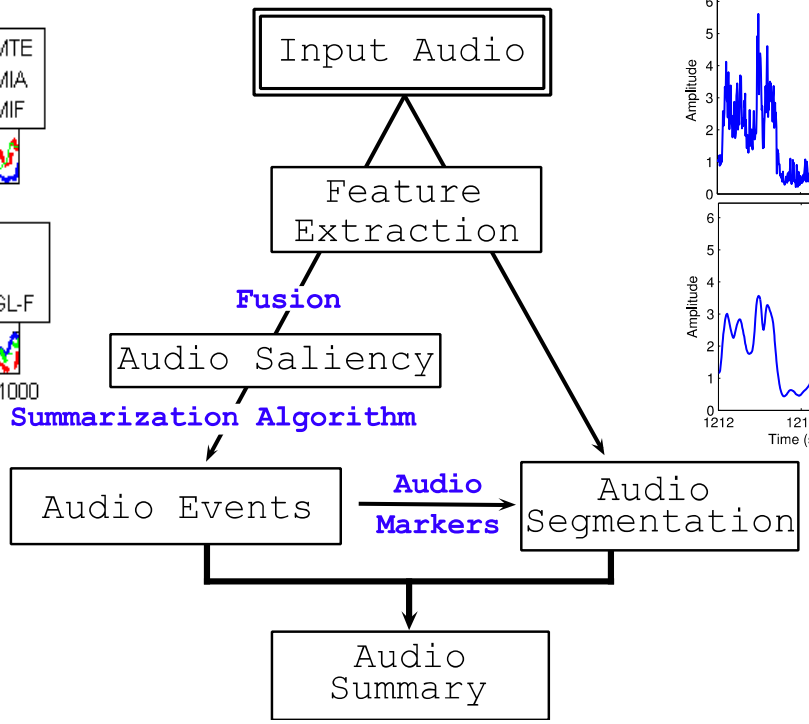
| AUC Results | V-V | | A-A | | AV-AV (mean) | |
|---|---|---|---|---|---|---|
| videos | Hndcr. | CNN | Hndcr. | CNN | Hndcr. | CNN |
| Full Length Movie | | | | | | |
| GWW* | 0.589 | 0.644 | 0.714 | 0.706 | 0.664 | **0.735** |
| GWW** | 0.626 | **0.660** | 0.706 | **0.740** | 0.648 | 0.710 |

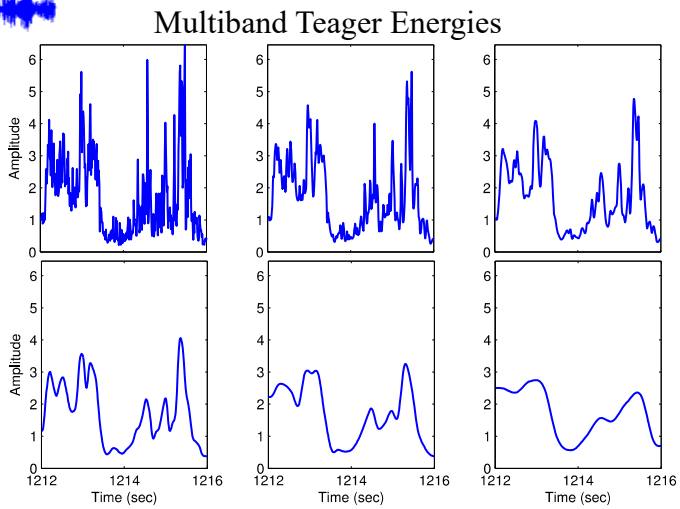
- For the “Gone with the Wind” movie two different setups were adopted:
  i. only the six Hollywood movies were used for training (GWW*)
  ii. all data was used for training (GWW**), thus six movies and five travel documentaries
- CNN-based architecture outperforms the hand-crafted frontend for all modalities
  - Better improvements when CNN models are trained in all data

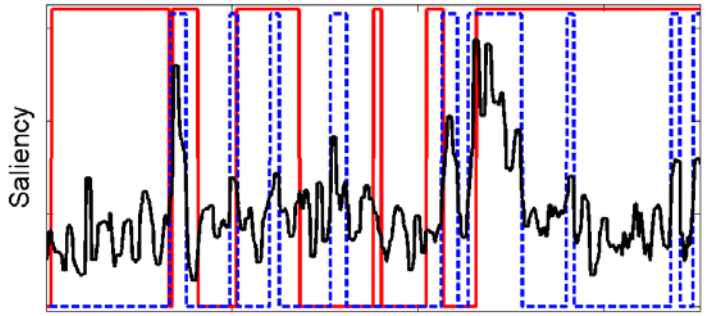
# Applications, Demos

# Audio Summarization System Overview




**Saliency Curve**
**Audio annotated salient segments**
**Summary x2: Included segments**

# Video Summaries

**AR London ca 16% ca 3’40”**


**GWTW ca 3% ca 3’**
(3min from full duration movie)

# Part 4: Conclusions

- Two approaches for audio-visual salient event detection:

i. Handcrafted frontend for multimodal saliency: improved synergistic approach based on a unified energy-based audio-visual framework

ii. CNN-based architectures

- Summarization method for the production of automatic summaries
- Evaluation on COGNIMUSE database
  - ❑ different types of videos and mono- and multimodal ground-truth annotations of the salient events
  - ❑ CNN-based architecture outperform almost in all cases the handcrafted frontend

➢ Multimodal (Audio-Visual) saliency outperforms the results of monomodal saliency in many cases

Tutorial slides: http://cvsp.cs.ntua.gr/interspeech2018
COGNIMUSE dataset: http://cognimuse.cs.ntua.gr/database
For more information, demos, and current results: http://cvsp.cs.ntua.gr and http://robotics.ntua.gr