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

Part 5
Multimodal Video Summarization

Petros Maragos, Alexandros Potamianos,
Athanasia Zlatintsi and Petros Koutras

Tutorial at IEEE International Conference on Acoustics, Speech and Signal Processing 2017,
New Orleans, USA, March 5, 2017

Part 5: Outline

- State-of-the-art in Video Summarization
- COGNIMUSE Database: Saliency, Semantic & Cross-Media Events Database
- Movie Summarization System #1 (Bottom-Up, Fusion)
- Movie Summarization System #2 (Improved Frontends, Learning)

Video Summarization

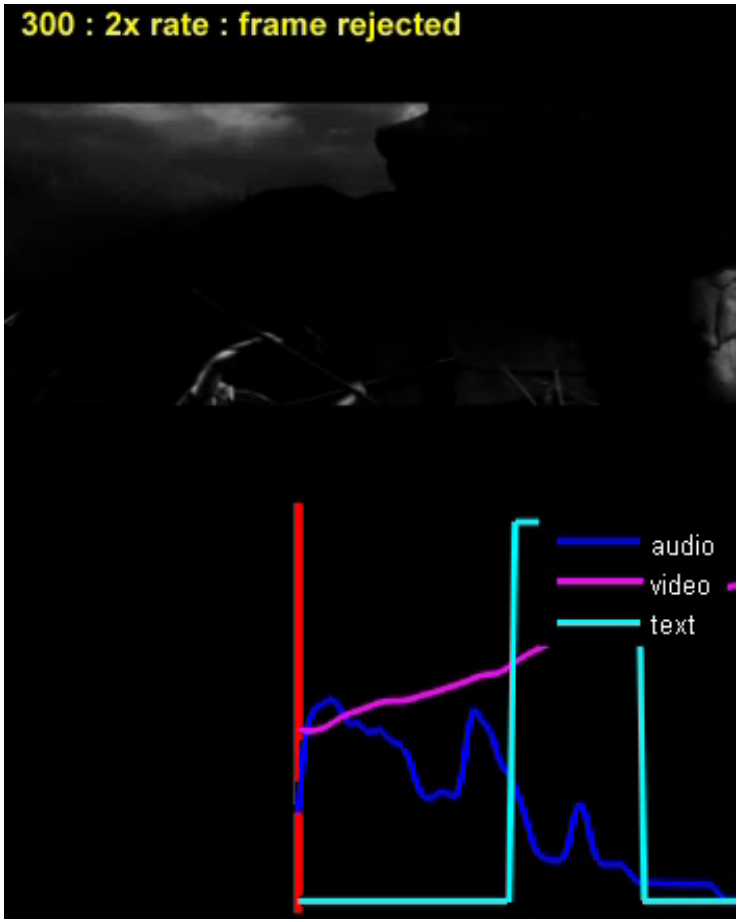
- Summarization task refers to producing a shorter version of a video:
  - containing all the necessary information required for context understanding
  - without sacrificing much of the original informativeness and enjoyability
- Automatic summaries can be created with:
  - key-frames, which correspond to the most important video frames and represent a static storyboard
  - video skims that include the most descriptive and informative video segments

Key-frame Summary

Original frames: 7414, Key-frames: 31, Skimming percentage: 0.42%

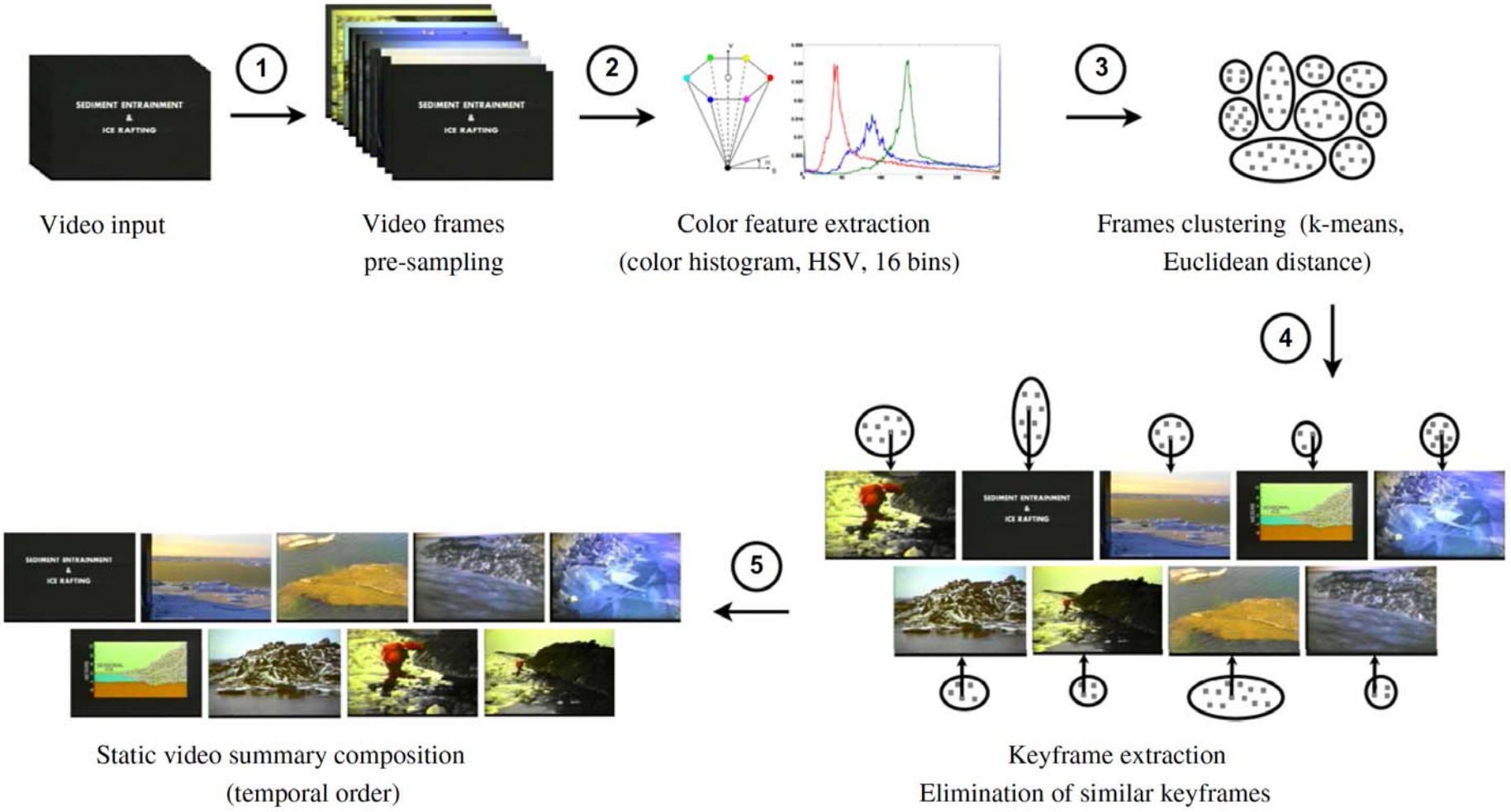
"Cold Mountain", Miramax Films, 2003.

Demo: Video Skimming Example with Accept/Reject Frames and A-V-T Saliency Curves



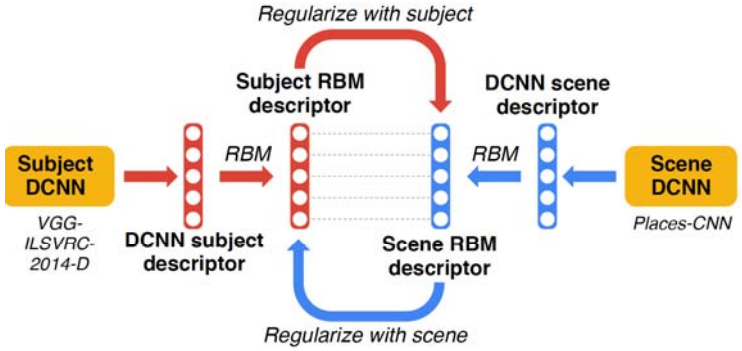
State-of-the-Art in Video Summarization

VSUMM: Static Summaries

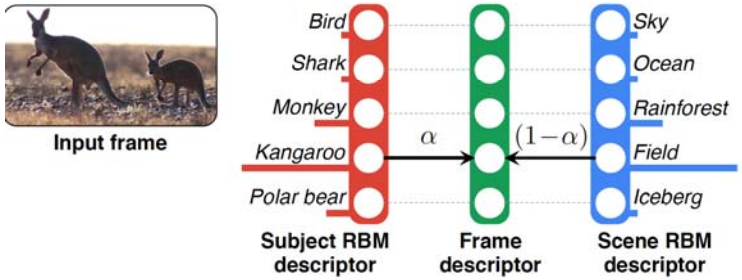


[S.E.F. De Avila, A.P.B. Lopes, A. da Luz and A. de Albuquerque Araújo, *VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method*, Pat. Rec. Let., 2011]

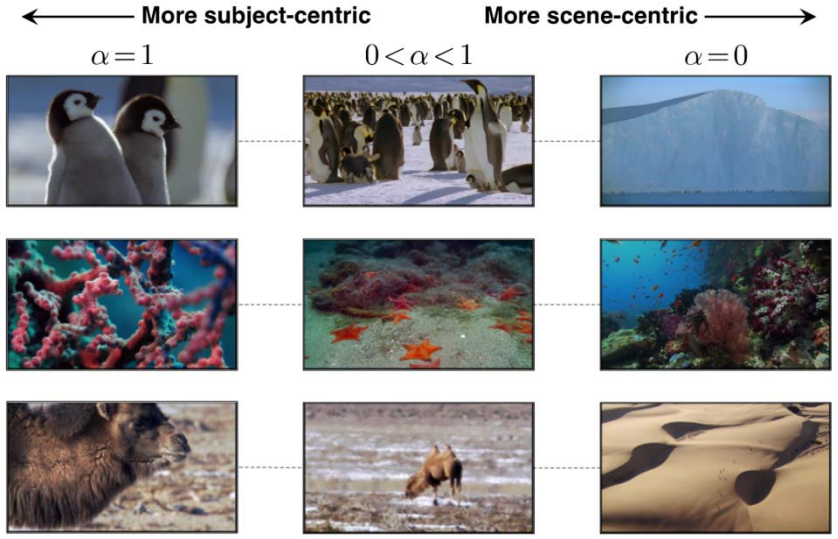
Co-Regularized Deep Representations

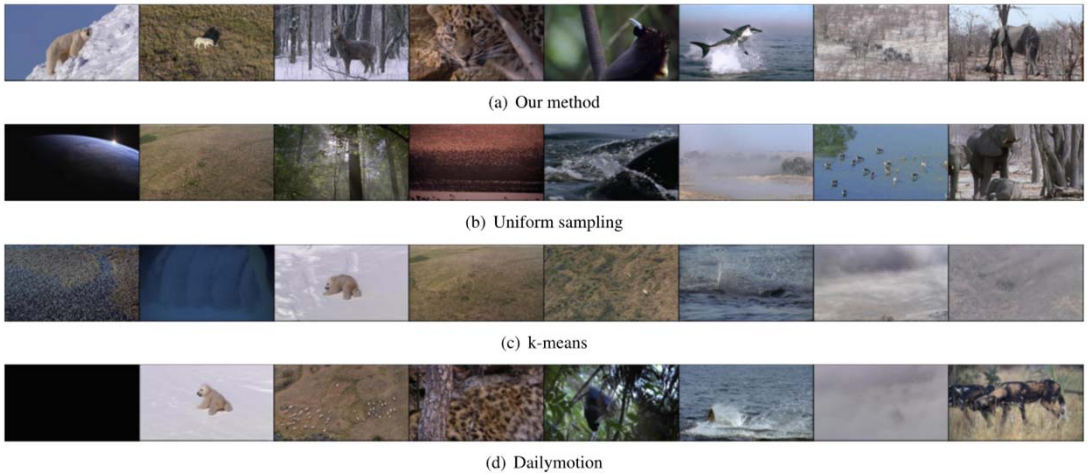


(a) Training co-regularized RBMs.



(b) Generating a frame descriptor from co-regularized RBMs.





(a) Our method

(b) Uniform sampling

(c) k-means

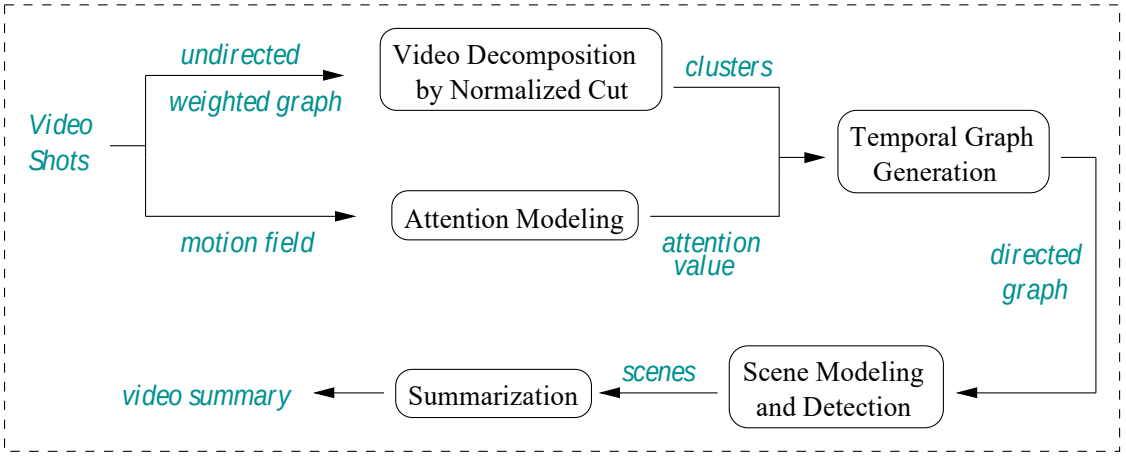
(d) Dailymotion

- Keyframe-based summarization framework combining deep convolutional neural networks (DCNNs) and restricted Boltzmann machines (RBMs)
- Co-regularization scheme for restricted RBMs able to learn joint high-level subject-scene representations

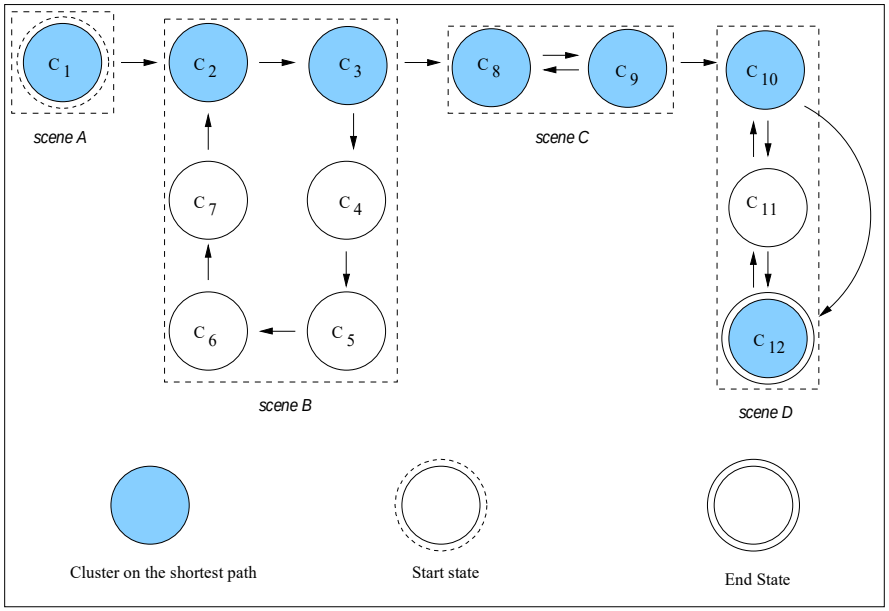
[O. Morere, H. Goh, A. Veillard, V. Chandrasekhar and J. Lin, *Co-Regularized Deep Representations for Video Summarization*, ICIP 2015]

Graph Modeling

- Summarization based on the analysis of video structures and highlights
- Normalized cut algorithm to globally and optimally partition a video into clusters



- Motion attention model based on human perception for computation of the perceptual quality of shots and clusters
- Cluster and computed attention form a temporal graph similar to Markov chain, describing the evolution and perceptual importance of video clusters



[C.-W. Ngo, Y.-F. Ma and H.-J Zhang, *Automatic Video Summarization by Graph modeling*, ICCV 2003]

Summarizing User Videos

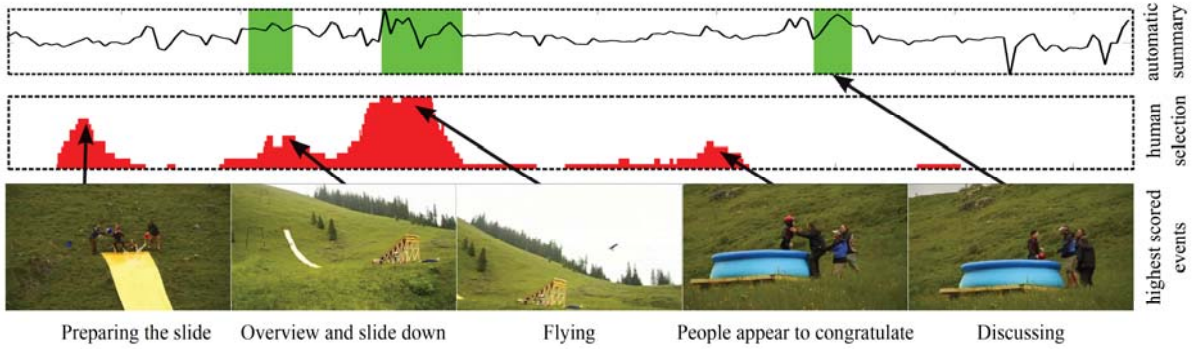
Summaries as optimization problem:

$$\underset{\mathbf{x}}{\text{maximize}} \sum_{i=1}^{n} x_i I(S_i)$$

$$\text{subject to} \sum_{i=1}^{n} x_i |S_i| \leq L_s$$

Per-frame Interestingness:

- Attention
- Aesthetics/Quality
- Presence of landmarks
- Faces/Persons
- Follow object

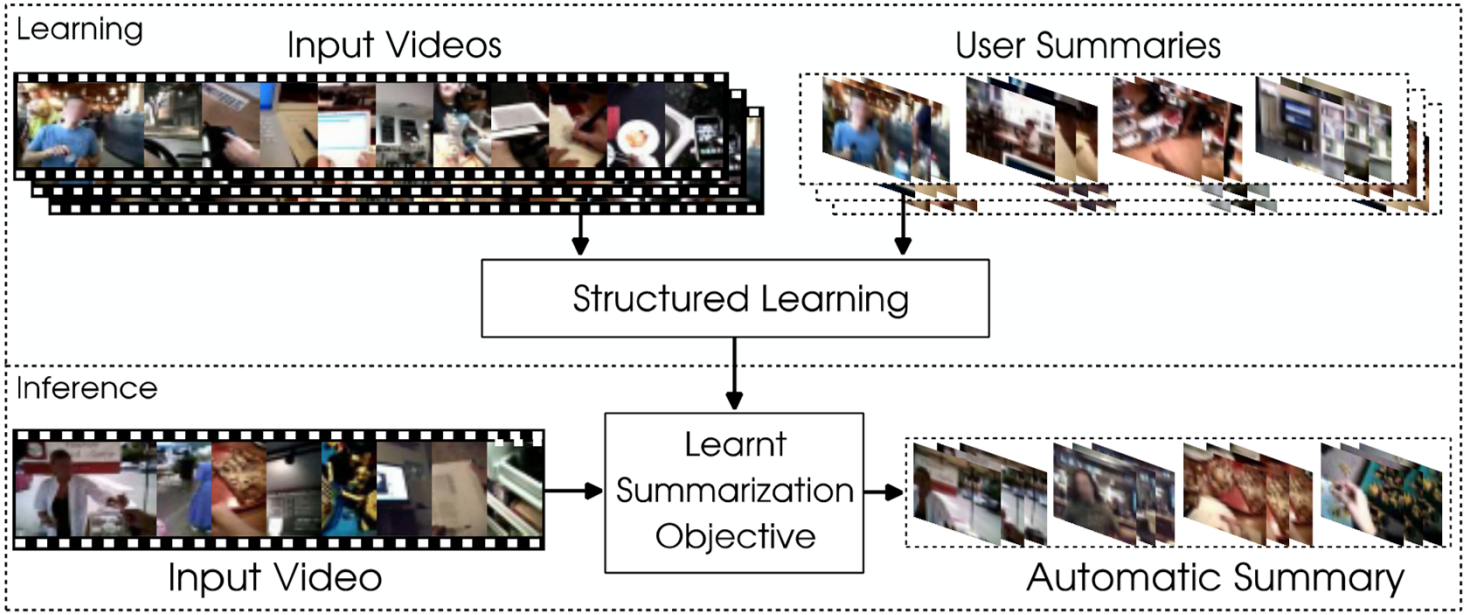


[M. Gygli, H. Grabner, H. Riemenschneider and L. Van Gool, "*Creating summaries from user videos*", ECCV 2014]

Learning Mixtures of Objectives

Submodular functions capturing the quality of summary:

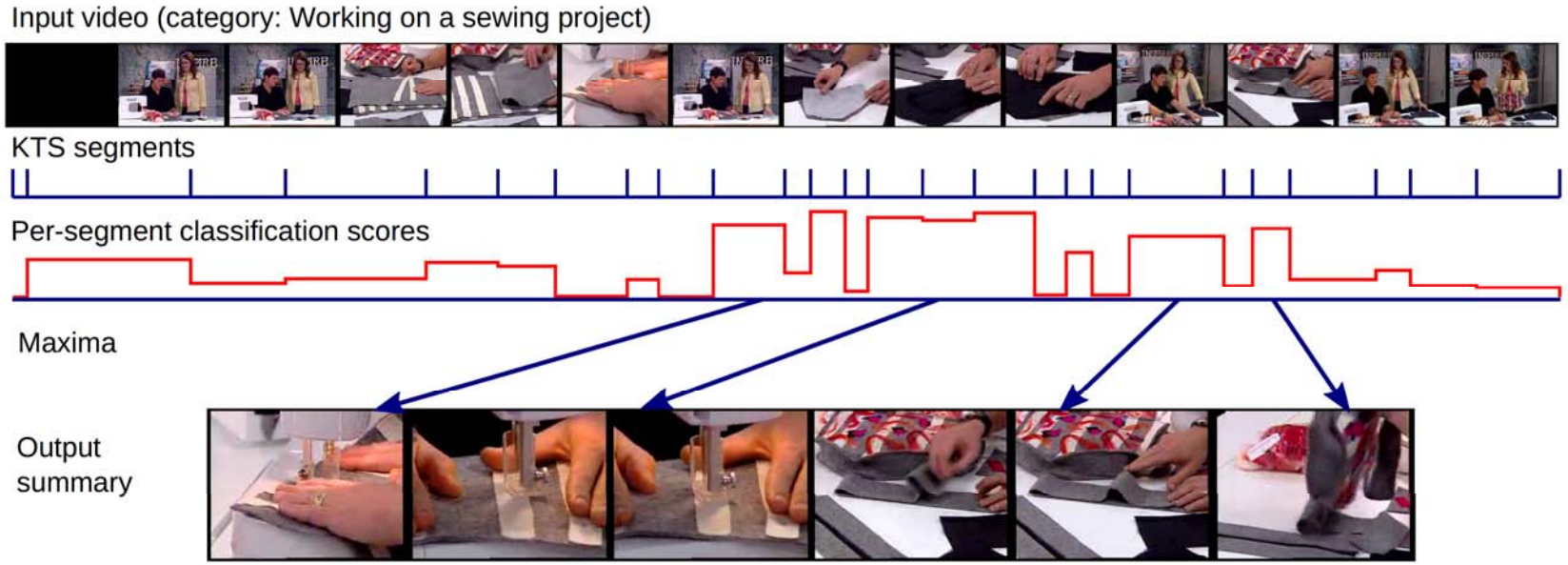
- Interestingness
- Representativeness
- Uniformity



[M. Gygli, H. Grabner and L. Van Gool, *Video summarization by learning submodular mixtures of objectives*, CVPR 2015]

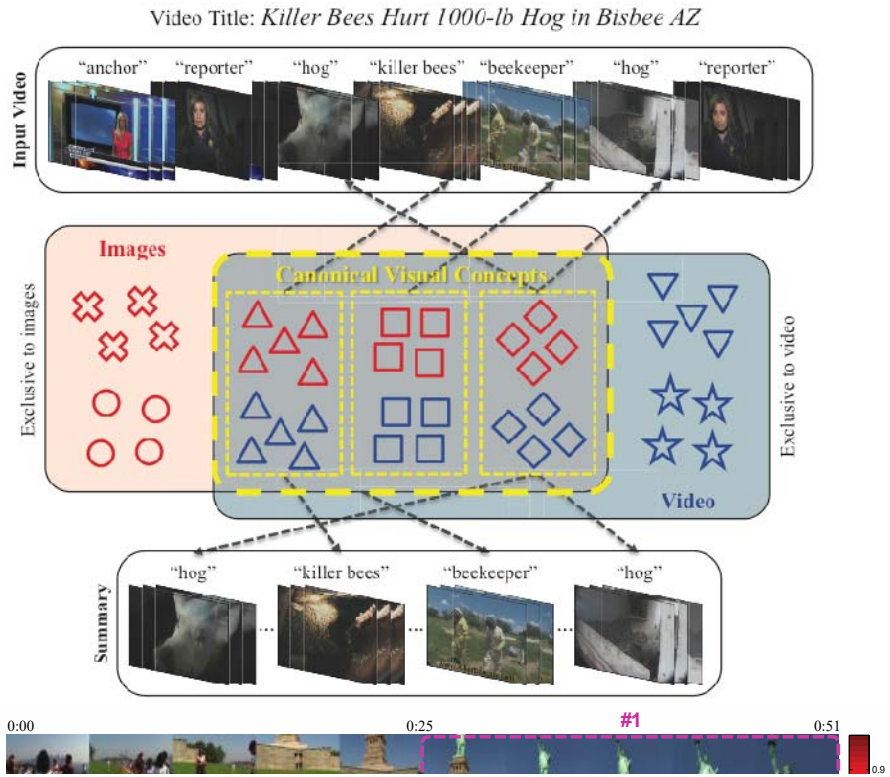
Category-specific Video Summarization

- Temporal Video Segmentation
  - Kernel Temporal Segmentation Algorithm (KTS)
- Supervised summarization
  - Train a linear SVM with just video-level class labels
  - Score segment descriptors with the classifiers scores



[D. Potapov, M. Douze, Z. Harchaoui and C. Schmid, "*Category-specific video summarization*", ECCV 2014]

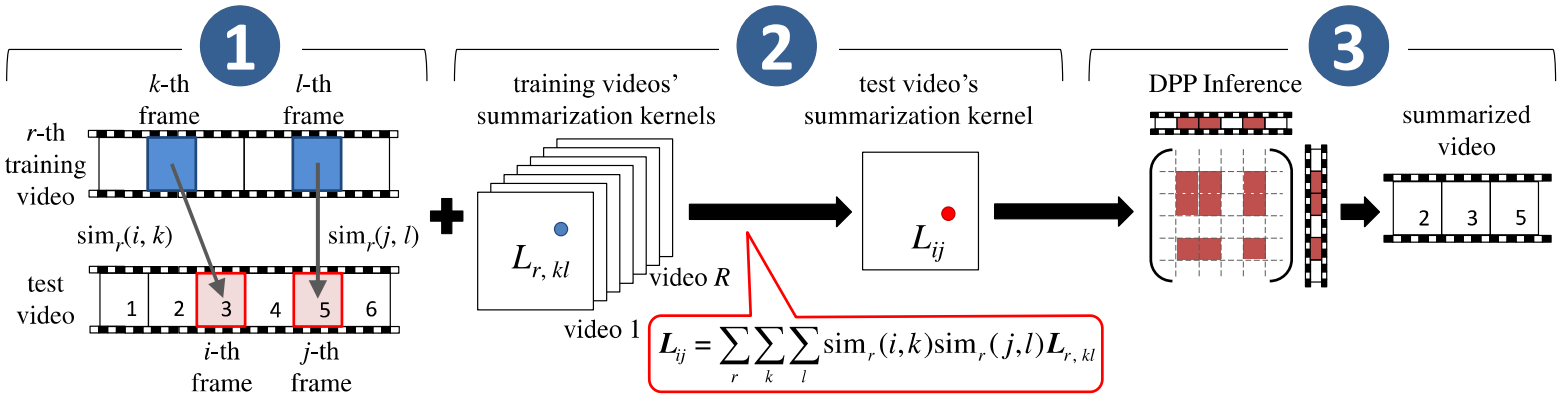
TVSum: Summarizing Web Videos Using Titles



- Title-based image search
- Generate a summary by selecting shots that are the most relevant to (representative of), canonical visual concepts shared between the given video and images
- Learn canonical visual concepts by focusing on the shared region (yellow dotted rectangle area), singling out patterns that are exclusive to either set
- Discard images irrelevant to video frames (and vice versa)

[Y. Song, J. Vallmitjana, A. Stent and A. Jaimes, *TVSUM: Summarizing Web Videos Using Titles,* CVPR 2015]

Summary Transfer

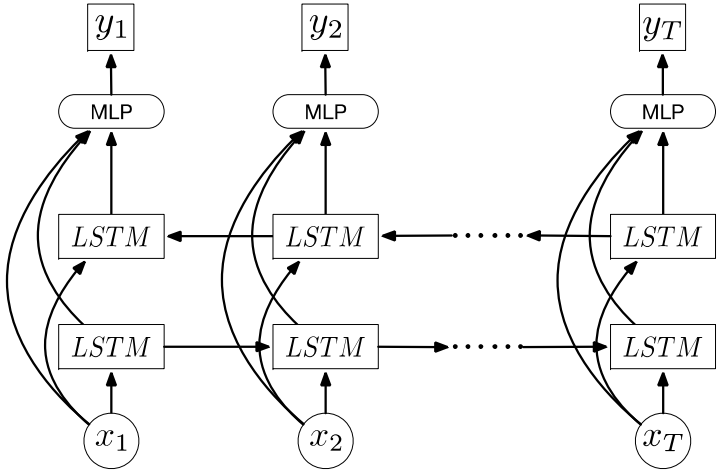


Similar videos share similar summary structure

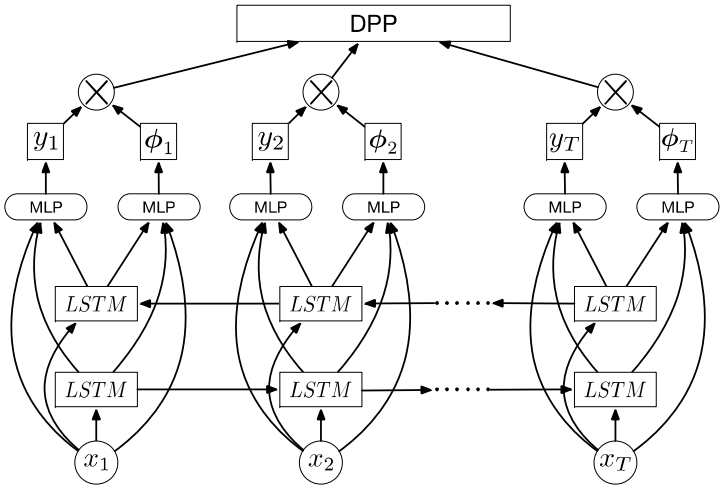
- Nonparametrically transfer summary structures from annotated videos to unseen test videos
- Compute frame-level similarity between training and test videos
- Encode summary structures in the training videos with kernel matrices made of binarized pairwise similarity among their frames
- Semantic side information about the video's genre
- Sub-shot based summarization

[K. Zhang, W.L. Chao, F. Sha and K. Grauman, *Summary Transfer: Exemplar-based Subset Selection for Video Summarization*, CVPR 2016]

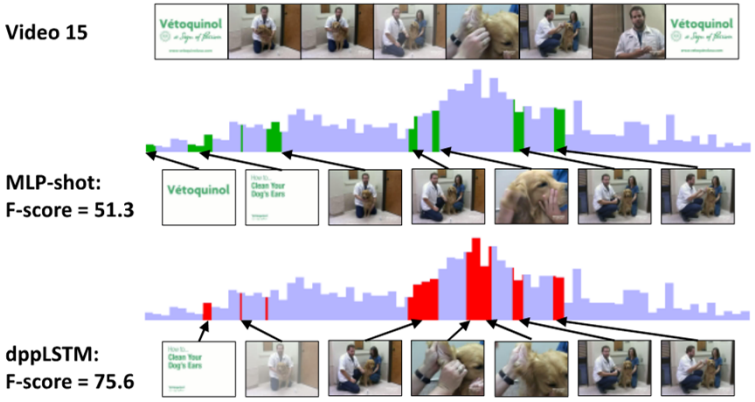
Video Summarization with Long Short-term Memory



vsLSTM model: composed of two LSTM layers: modeling video sequences in the forward direction and the other the backward.



dppLSTM model: combining vsLSTM and DPP by modeling both long-range dependencies and pairwise frame-level repulsiveness explicitly

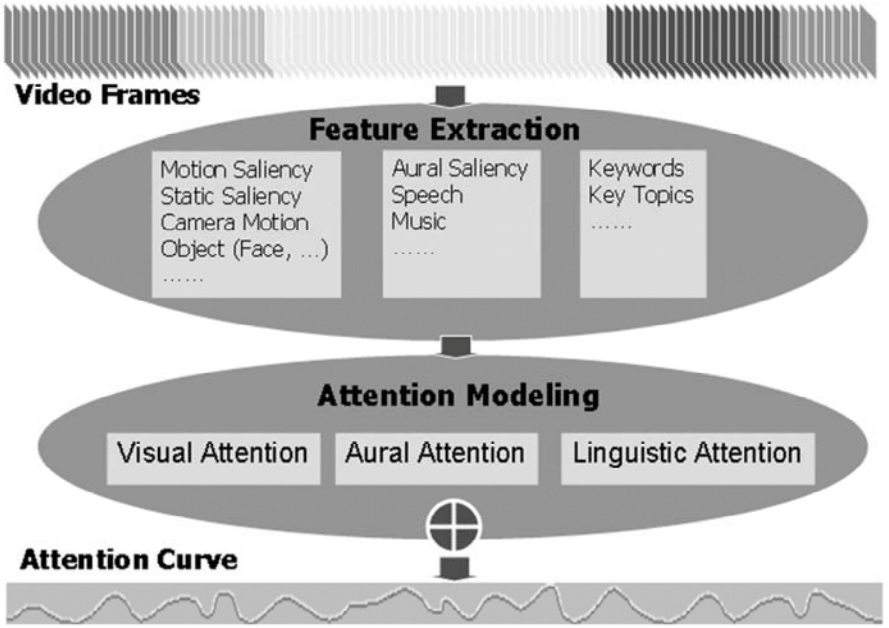


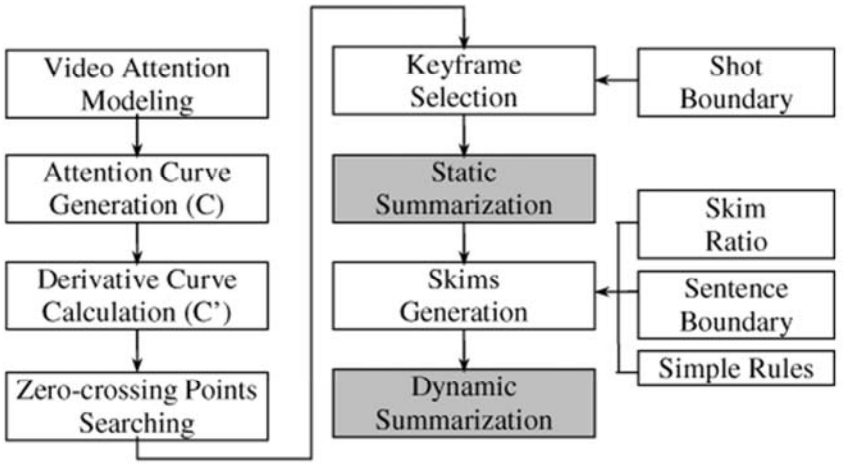
Example video summaries by a multi-layer perceptron (MLP-Shot) and dppLSTM, along with ground truth

[K. Zhang, W.-L. Chao, F. Sha and K. Grauman, *Video Summarization with Long Short-term Memory*, ECCV 2016]

Multimodal Attention Model for Summarization

- Visual Attention Modeling
  - Motion Attention Model
  - Static Attention Model
  - Face Attention Model
  - Camera Motion Model
- Aural Attention Modeling
  - Aural Saliency Model
  - Speech and Music Attention Models
- Fusion Schemes
  - Linear
  - Non-Linear





[Y.F. Ma, X.S. Hua, L. Lu and H.J. Zhang, *A generic framework of user attention model and its application in video summarization*, IEEE Trans. MM., 2005]

COGNIMUSE Database

Saliency, Semantic &
Cross-Media Events Database

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

[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., under review]
[A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos and K. Pastra, *Quality Evaluation of Computational Models for Movie Summarization,* QoMEX 2015]

COGNIMUSE Database
Saliency, Semantic & Cross-Media Events Database

Including:

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

Use: training and evaluation of event detection and summarization algorithms, classification/recognition of audio-visual events, emotion tracking

- Baseline experiments/results with state-of-the-art algorithms on all videos using MovieSum #2 System

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)

- Saliency annotated (all database videos) by **3 annotators** in separate runs for each individual layer
- Emotion annotation for the 7 movie clips
- Event annotation for the 7 movie clips
- COSMOROE annotation for GWTW
- Part-of-speech tags, sentence and word boundaries for the whole database

Database Annotation: Saliency & Structure

Movie Structure: Shots 370-699 (~540/movie) and 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….

Percentage (%) Salient Frames

Percentage (%) of Salient Frames on movies (labeled by at least two annotators)								
Layer	**BMI**	**CHI**	**CRA**	**DEP**	**GLA**	**LOR**	**FNE**	**GWW**
A	25.4	57.8	56.3	30.0	60.0	58.6	53.5	69.2
V	30.1	46.9	34.9	28.6	36.6	34.8	34.0	71.45
AV	27.4	46.8	39.8	36.2	48.9	39.9	38.8	70.06
AVS	63.2	76.6	64.8	71.8	68.5	64.8	67.9	88.0

Percentage (%) of Salient Frames on travel doc (labeled by at least two annotators)					
Layer	**AR_London**	**AR_Rio**	**AR_Tokyo**	**AR_Sydney**	**GoT_London**
A	58.7	43.8	60.3	55.1	45.6
V	49.5	48.5	46.6	48.8	40.5
AV	53.9	50.3	54.7	53.7	42.5
AVS	72.7	79.4	80.3	80.4	72.5

Event Annotation:
Multilayer Audio Event & Visual Action Annotation

Audio event annotation based on the classes of:

"**Urban Sounds Dataset**" - with additions when a specific event found in regular bases in a movie

Visual action annotation based on the classes of:

- **Hollywood2**
- **Bojanowski et al**. from ECCV 2014: 4 more actions added to Hollywood
- **HMDB:** a large human motion database

Urban Sounds: [J. Salomon, C. Jacoby and J.P. Bello, *A Dataset and Taxonomy for Urban Sound Research*, ACMMM 2014]
Hollywood 2: [M. Marszalek, I. Laptev and C. Schmid, *Actions in Context*, CVPR 2009]
HMDB: [Kuehne et al. *A Large Video Database for Human Motion Recognition*, ICCV 2011]
[P. Bojanowski et al., *Weakly Supervised Action Labeling in Videos Under Ordering Constraints*, ECCV 2014]

Categories of Acoustic Events

Categories	Layer 1	Layer 2
Human	Voice (x3)	speech_male, speech_female, speech_child, speech_synthetic, crowd_noise, laughter, shouting, crying, coughing, sneezing, breathing, spitting, singing, infant, other
	Movement (x3)	footsteps, punching, other
Nature	Elements (x2)	wind, water, waves, thunder, fire, sand, other
	Animals (x2)	dog_bark, dog_howl, bird_tweet, bird_sing, horse_galloping, horse_neighing, sheep, other
	Plants/Vegetation (x2)	leaves_rustling, other
Mechanical	Construction (x2)	jackhammer, hammering, drilling, sawing, explosion, engine_running, other
	Ventilation (x2)	air-conditioner, other
	Non-motorized Transport (x2)	bicycle, skateboard, other
	Social/signals (x2)	bells, clock chimes, alarm/siren, fireworks, gun shot, explosion, glass_breaking, door rusty, door_opening/closing, swords, other
	Motorized Transport (x2)	marine, rail, road, air, other
Music	Non-amplified (x1)	live
	Amplified (x1)	live, recorded
	Sound Source (x1)	Diegetic: originated from the source within the film's world, Non-diegetic: mood music Background music: when music is not the basic element in the scene Foreground music: when music is basically the only thing you hear
	Genre (x1)	classical, symphonic, rock, pop, punk, jazz, folk/country, blues, metal, rock 'n roll, hiphop, reggae, electronic, funk/soul/rnb, ethnic/world, other
	Instrument (x1)	keyboard, string, wind, percussion, orchestra, electronic/amplified, mixed (e.g., rock band etc.), other

Types of visual events

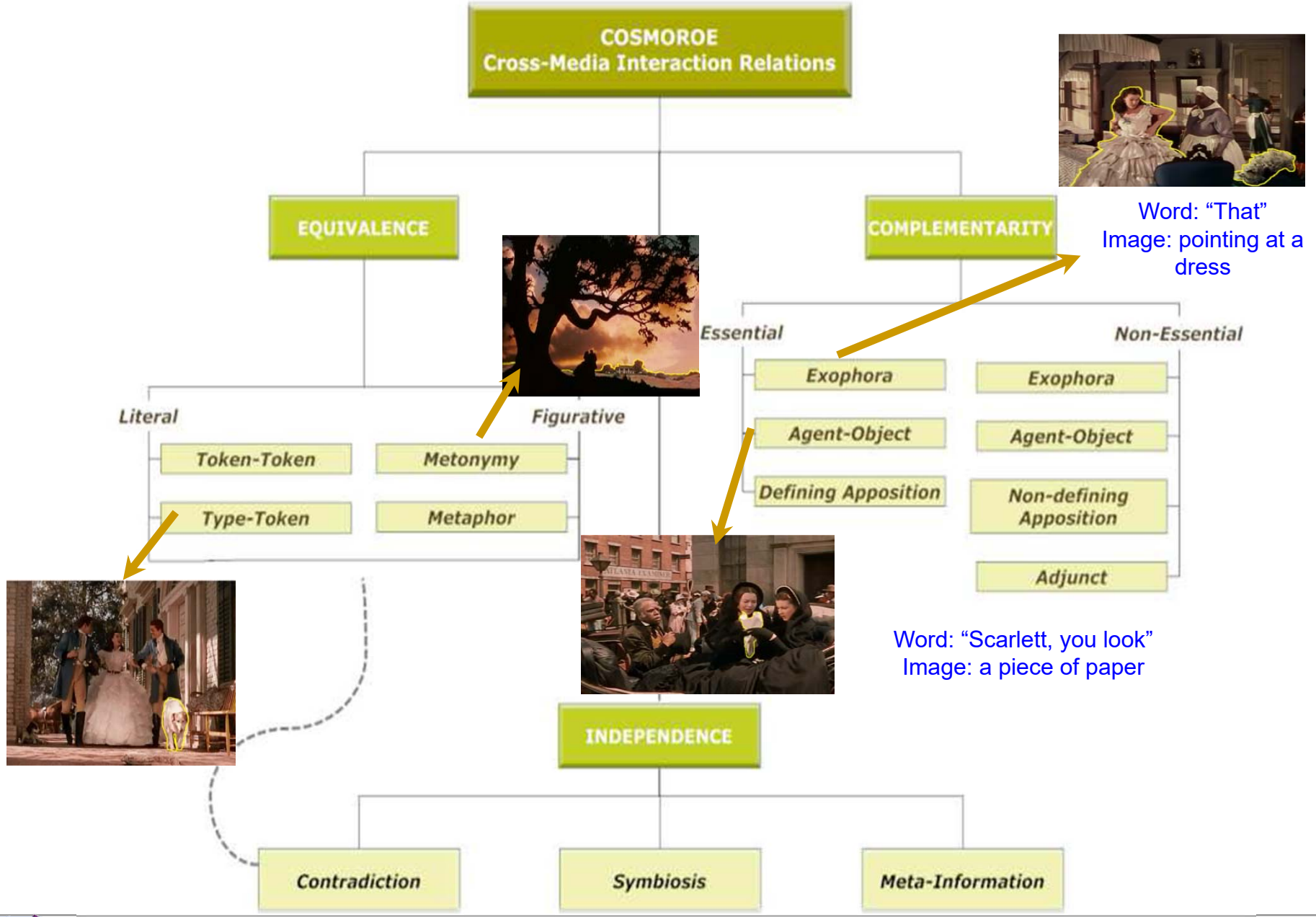
Categories	Layer 1
General facial actions (x2)	smile, cry, laugh, chew, talk, other
Facial actions with object manipulation (x2)	smoke, eat, drink, other
General body movements (x2)	sitting down, sitting up, standing up, running, cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, somersault, turn, dance, walk, other
Gestures (x2)	wave hands, point something, pantomime, other
Body movements with object interaction (x2)	answering phone, driving car, getting out of the car, open car door, open door, brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick , pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw, Other
Body movements for human interaction (x2)	fighting, hugging, kissing, grab hand, threaten person, fencing, kick someone, punch, shake hands, sword fight, other

Statistics for Audio and Visual Events

Audio Events # instances: 6262, Total Duration in Hours: 19.24		
Category/Subcategory 1	Instances	Dur. (min.)
Voice	3809	245.75
Movement	228	19.82
Elements	154	16.91
Animals	222	20.26
Plants	0	0.00
Construction	46	5.19
Ventilation	4	0.54
Non-motorized Trans.	18	0.84
Social Signals	444	15.66
Motorized Trans.	48	3.86
Non-Amp. Music	12	5.16
Amplified	218	213.28
Sound Source	640	226.91
Genre	231	222.86
Intstrument	200	162.80

Visual Actions # instances: 4847, Total Duration in Hours: 4.58		
Category	Instances	Dur. (min.)
General facial actions	2233	129.67
Facial action with obj. manip.	90	4.08
General body mov.	1215	79.75
Gestures	284	9.09
Body mov. with object inter.	693	33.72
Body mov. for human inter.	332	18.79

Most Frequent Audio and Visual Events		
Category/Subcategory	Instances	Dur. (min.)
Voice: speech male	1874	102.39
Voice: speech female	1048	55.55
Voice: crowd noise	188	42.68
Sound source: background music	350	158.20
Sound source: foreground music	290	68.71
Genre: symphonic	119	118.61
Genre: other genre	70	42.14
Instrument: string	32	23.29
Instrument: percussion	102	91.86
Instrument: mixed	16	22.71
General Facial Actions: talk	1915	114.67
General Body Mov.: walk	456	41.72


COSMOROE
Cross-Media Interaction Relations
EQUIVALENCE
COMPLEMENTARITY
Literal
Figurative
Token-Token
Type-Token
Metonymy
Metaphor
Essential
Non-Essential
Exophora
Agent-Object
Defining Apposition
Exophora
Agent-Object
Non-defining Apposition
Adjunct
INDEPENDENCE
Contradiction
Symbiosis
Meta-Information
Word: "That"
Image: pointing at a dress
Word: "Scarlett, you look"
Image: a piece of paper

Emotion Annotation

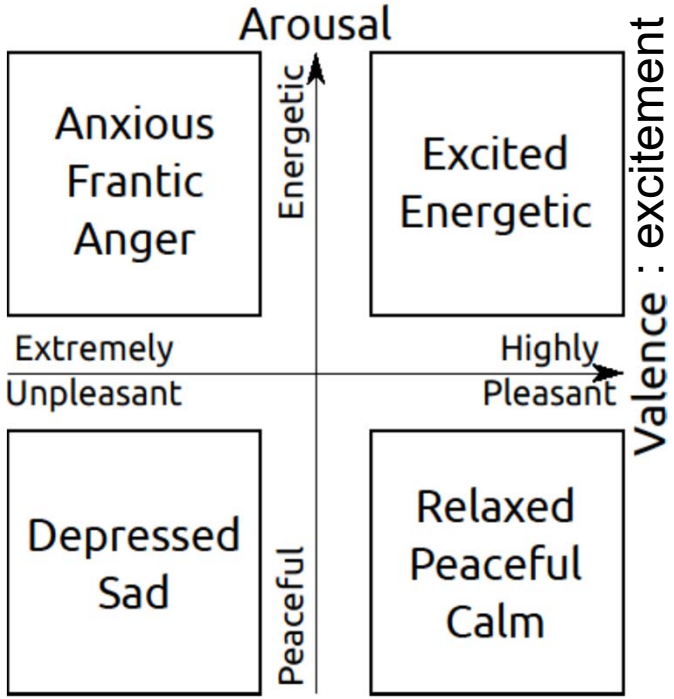
Intended emotion: (1 expert)

Annotations are meant to capture the emotional response that the movie tries to evoke in the viewer

Experienced emotion: (7 students)

annotations represent the actual emotional experience of an individual while watching a movie

Two time-series in [-1,1] for arousal and valence

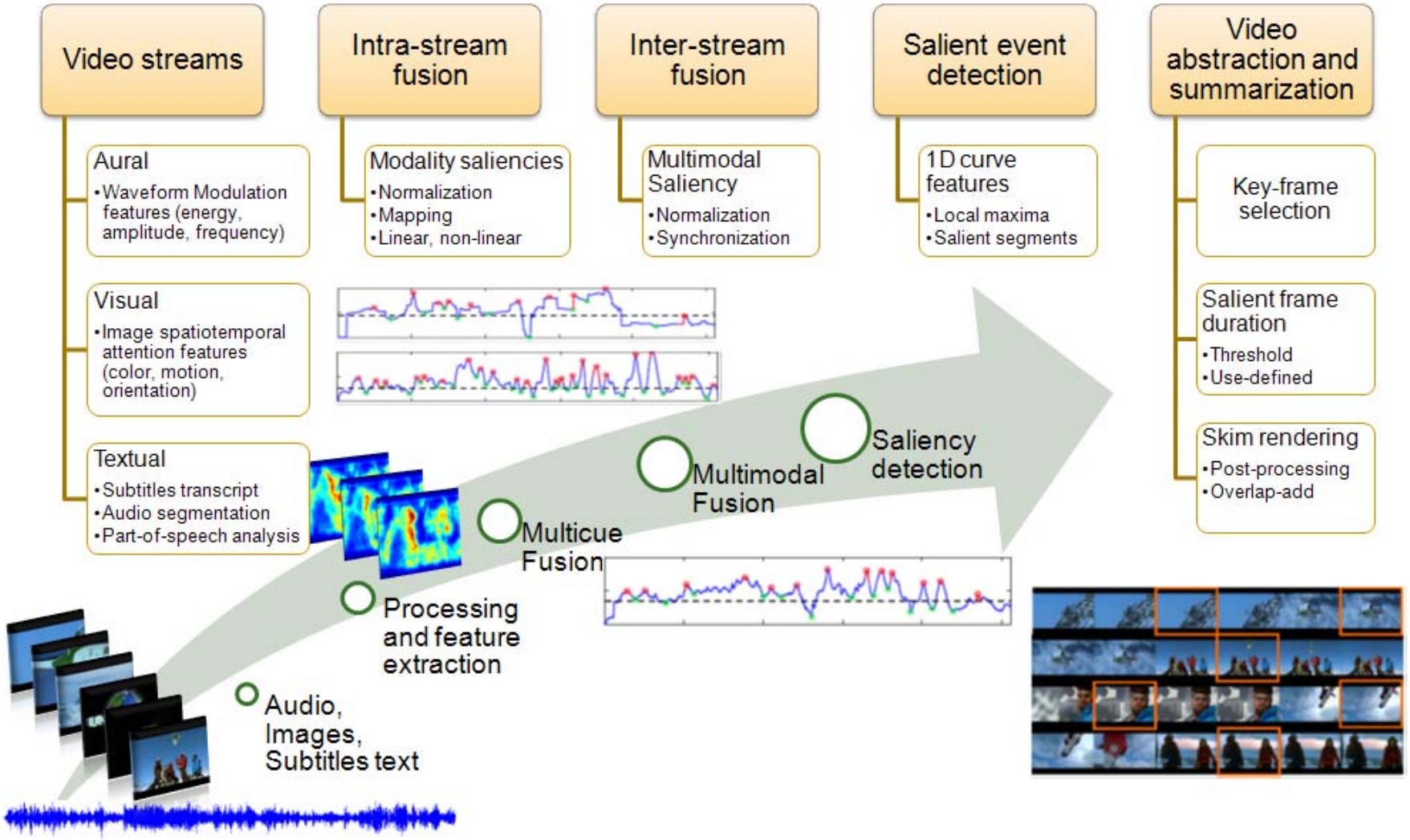


[N. Malandrakis, A. Potamianos, G. Evangelopoulos and A. Zlatintsi, *A Supervised Approach to Movie Emotion Tracking*, ICASSP 2011]

Movie Summarization System #1 (Bottom-Up, Fusion)

[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.]

Summarization System Overview (Fusion)



Audio Analysis: Saliency and Fusion

- Audio saliency cues
  - extracted through nonlinear operators

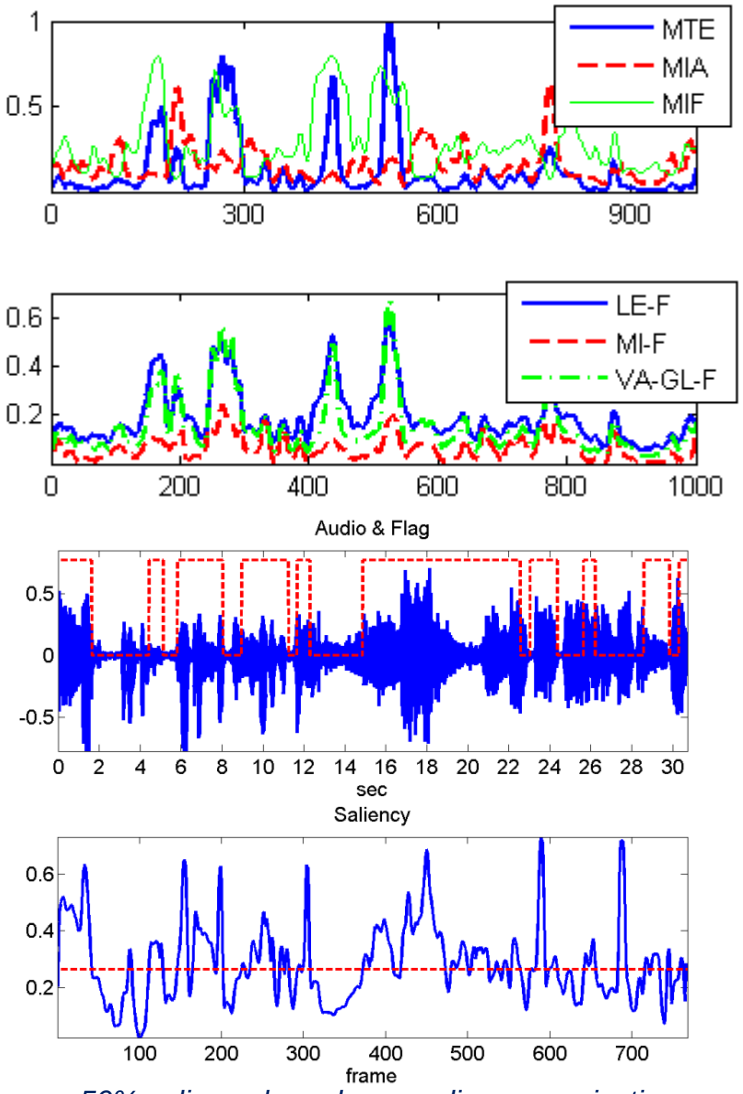
  Convey information on:
  - excitation level
  - frequency content
  - source energy tracking

- 3D Feature vector formation

$$\vec{F}_{\mathrm{a}}[m]=(\mathrm{MTE},\mathrm{MIA},\mathrm{MIF})[m]$$

- Monomodal saliency curve ()
  - Continuous-valued indicator of salient events, in [0,1]

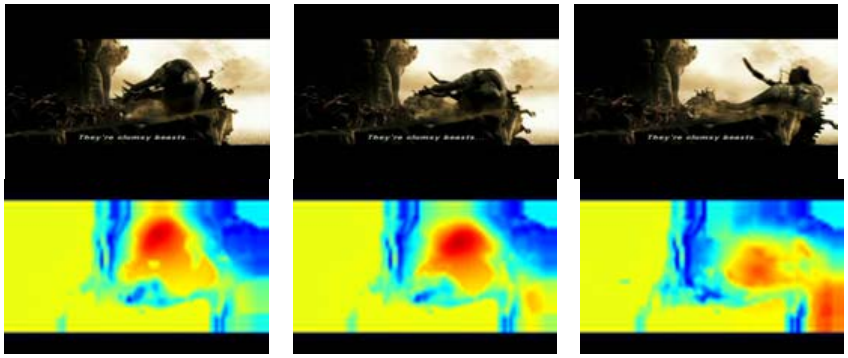
$$S_A = \mathrm{fusion}(S_1, S_2, S_3)$$



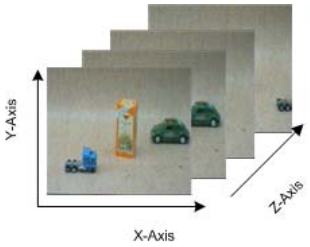
50% saliency-based raw audio summarization

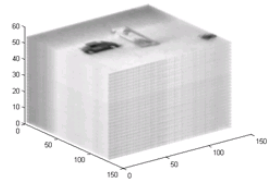
Visual Analysis: Features → Saliency Curve

- Spatio-temporal attention
  - Saliency maps

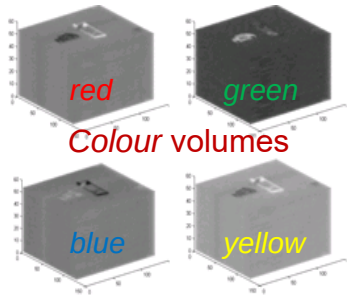


- Visual features
  - intensity
  - color (opponent model)
  - spatiotemporal orientations
    - steerable filters, orientation fusion

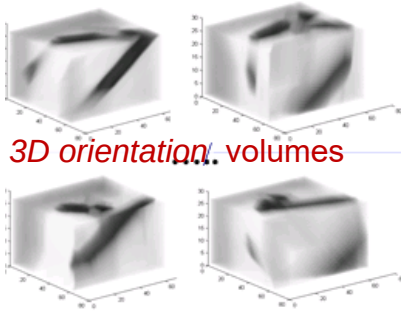




Intensity volume

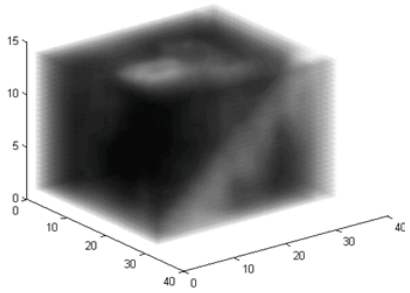


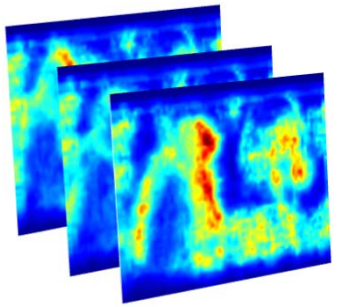
Colour volumes

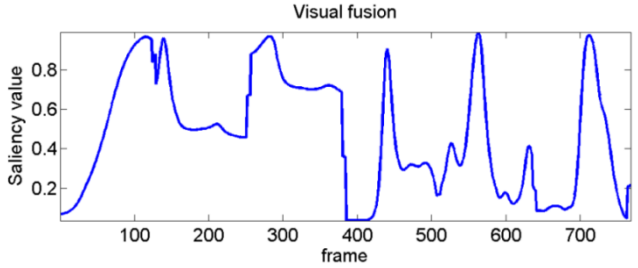


3D orientation volumes

- Pyramid, feature volumes, Saliency Volume → Saliency Visual Curve







Text Processing and Saliency

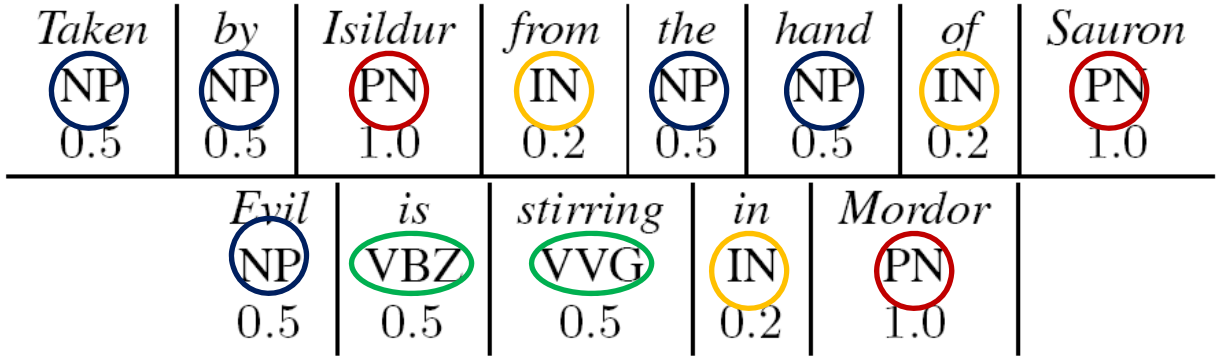
Movie Subtitles provide Text, Timestamps, Semantics.

Easy to process:

I. Extract movie transcript from subtitles and perform part-of-speech (POS) tagging

II. Segment audio stream using automatic speech recognition & forced alignment, and find the beginning/ending frame for each word in the transcript

III. Assign text saliency value to video frames based on parser tags



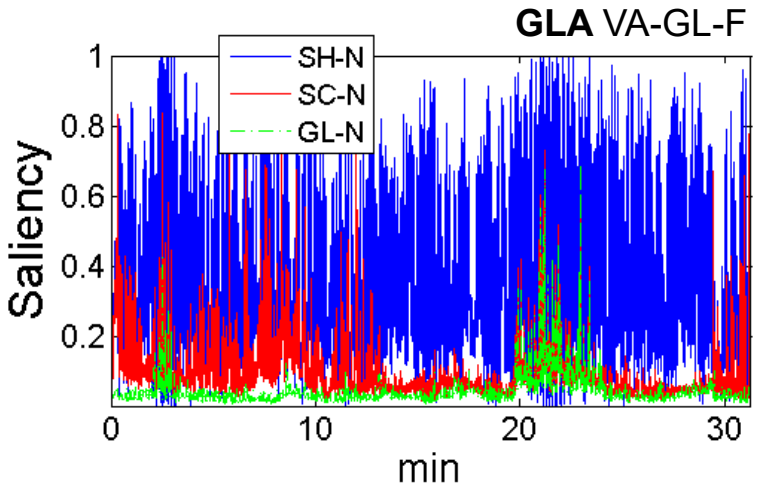
Fusion I: Scalar Operations

- Nine Fusion schemes $S_A = \text{fusion}(S_1, S_2, S_3)$
  - **Linear** (equal weights) (Low-level, memoryless) $S_{\text{LIN}} = w_1 S_1 + w_2 S_2 + w_3 S_3$
  - **Variance-based** (adaptive weights) $S_{VAR} = \sum_i \left(\frac{S_i}{\text{var}(S_i)} \right) / \sum_i \left(\frac{1}{\text{var}(S_i)} \right)$
  - **Nonlinear**
    - MIN $S_{MIN} = \min\{S_1, S_2, S_3\}$
    - MAX $S_{MAX} = \max\{S_1, S_2, S_3\}$
    - Weighted MIN $S_{MIVA} = \min(S_1 - w_1, S_2 - w_2, S_3 - w_3\} + \max(w_1, w_2, w_3)$

$$\text{where } w_i = \log\left(\frac{1}{\text{var}(S_i)} \right)$$

Fusion II: Normalization

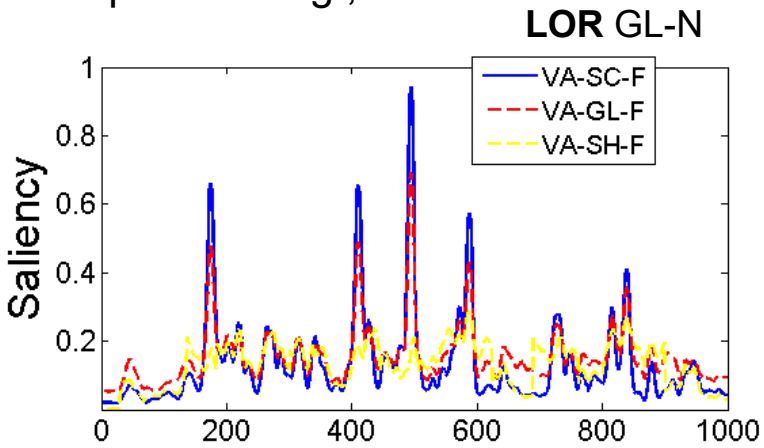
- Normalization intervals
  - Global linear normalization (GL)
  - Scene-based linear normalization (SC)
  - Shot-based linear normalization (SH)



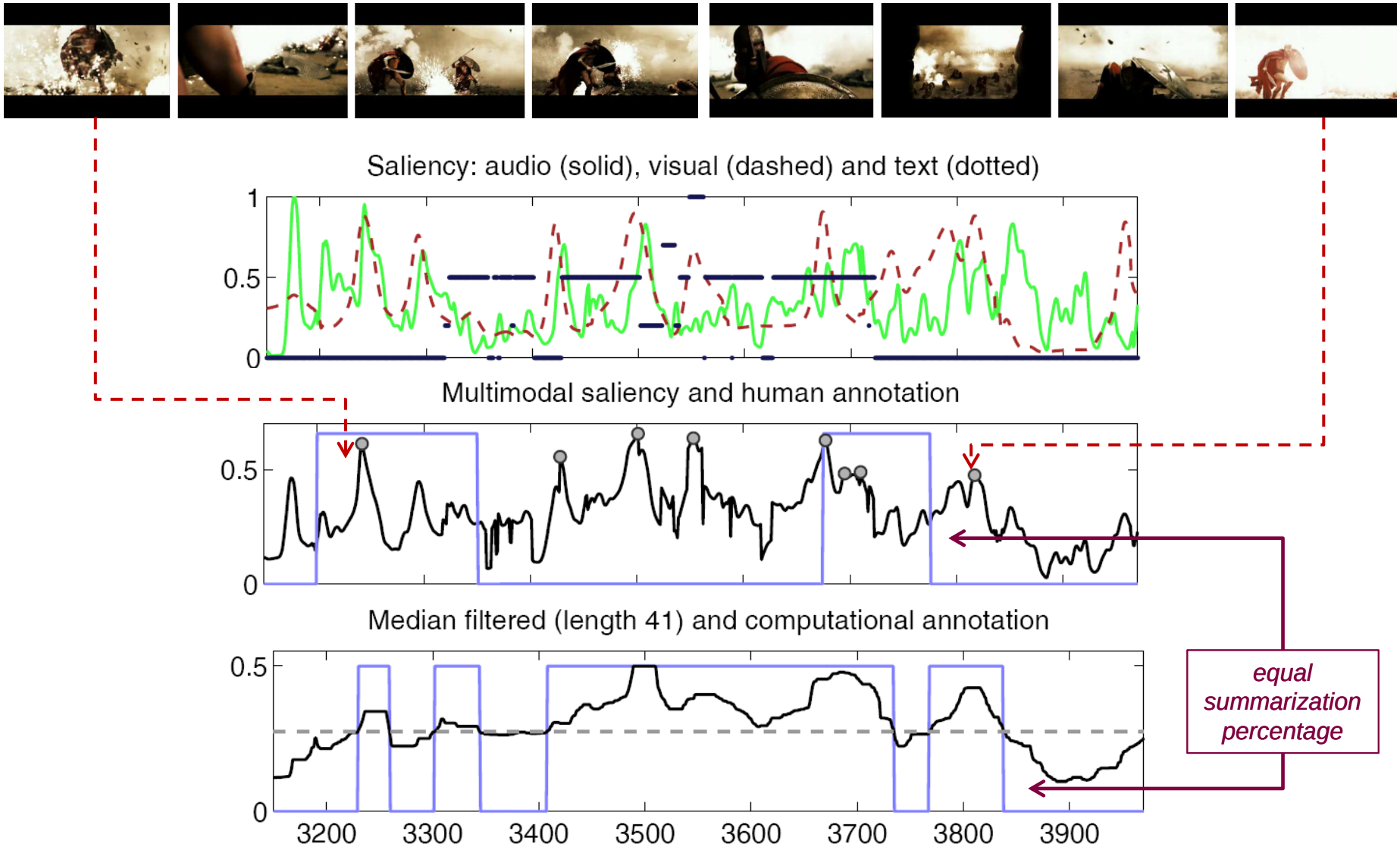
- Dynamic Adaptation levels

  i.e., weight updating with respect to Global or Local windows

  Inverse Variance & Weighted Min fusion can be computed at e.g.,
  - Global level (VA-GL)
  - Scene level (VA-SC)
  - Shot level (VA-SH)

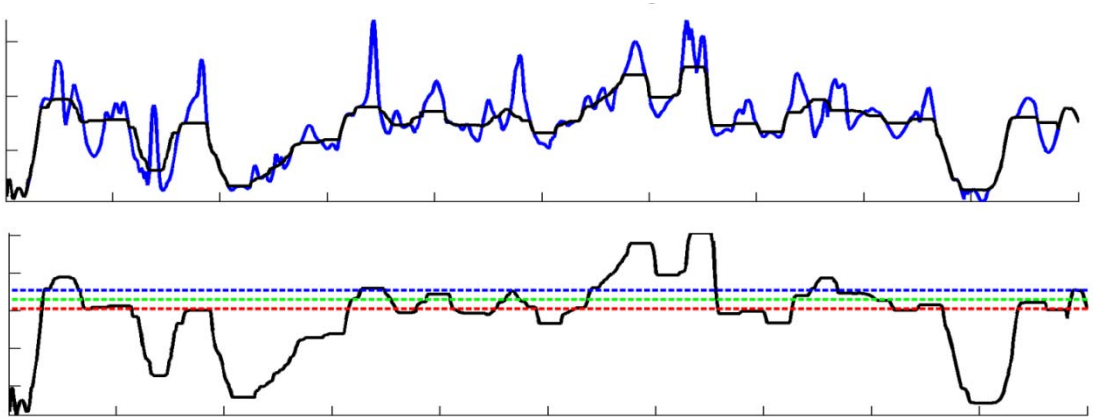


Multimodal Fusion: Audio, Visual, Text



Summarization Algorithm

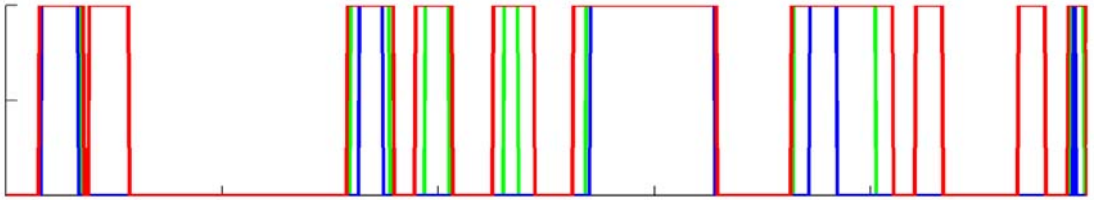
- Median-filtered Saliency (length $2M + 1$)
- Threshold selection
- Segment selection
- Reject: segments shorter than N frames
- Join: segments less than K frames apart
- Render: Linear overlap-add on L video frames and audio

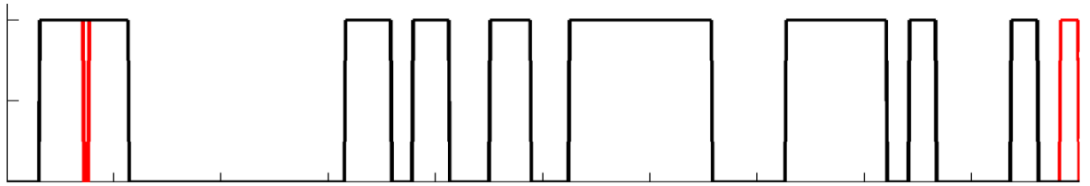


$$S_c : \text{card}\{D\} = c \cdot \text{length}\{\text{AVT}\},\ D = \{n : S_{\text{avt}}(n) > S_c\}$$

--- c = 0.2
--- c = 0.3
--- c = 0.5

$$S_{\text{avt}}(n) > S_c$$



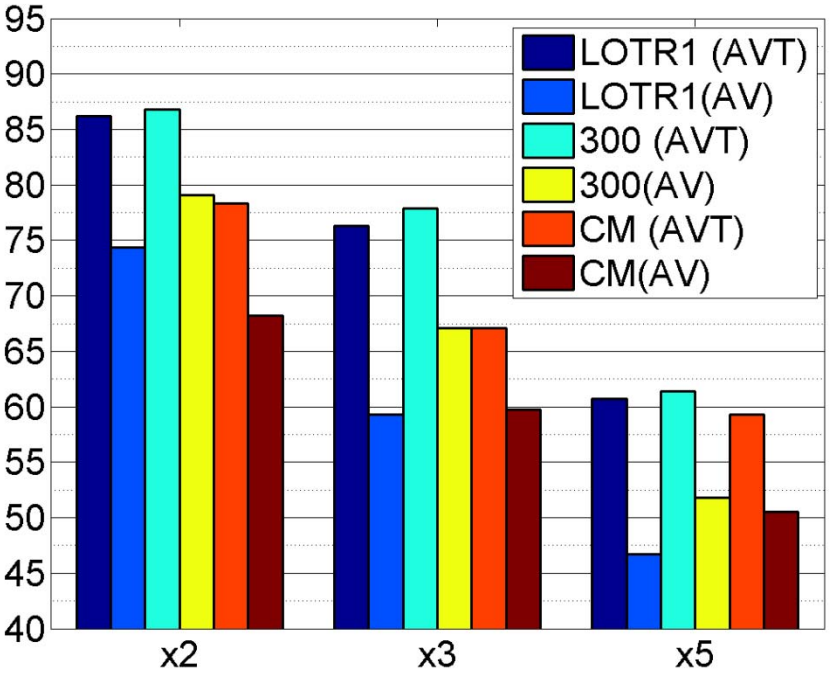


$$= (I_c \circ [1,\ldots,1]_N) \bullet [1,\ldots,1]_K$$

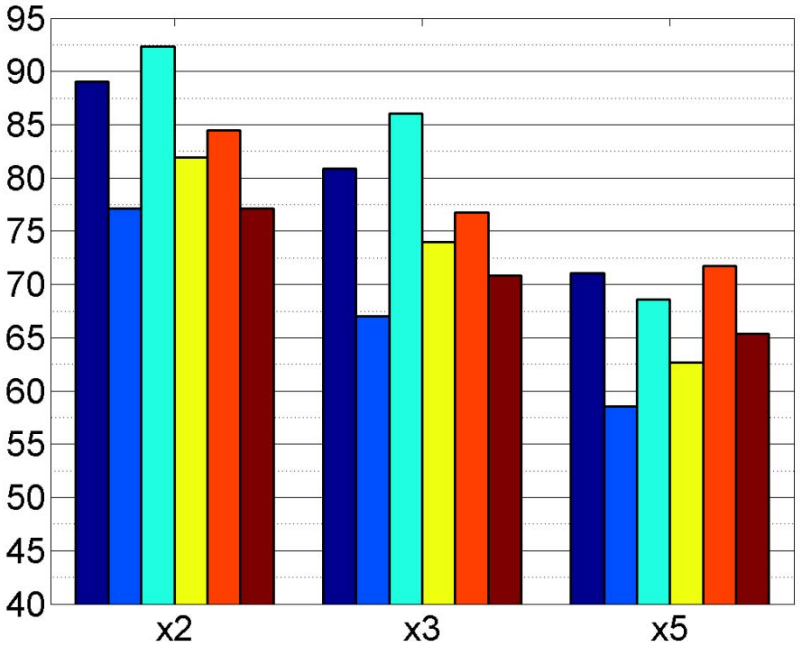
Subjective Results & Comparisons: AV to AVT

(Old MUSCLE database, 3 movies)

Evaluation by ~10 humans



Informativeness



Enjoyability

Informativeness x2:

AV only (without semantics) =~ 70+

AVT (with semantics) =~ 80+

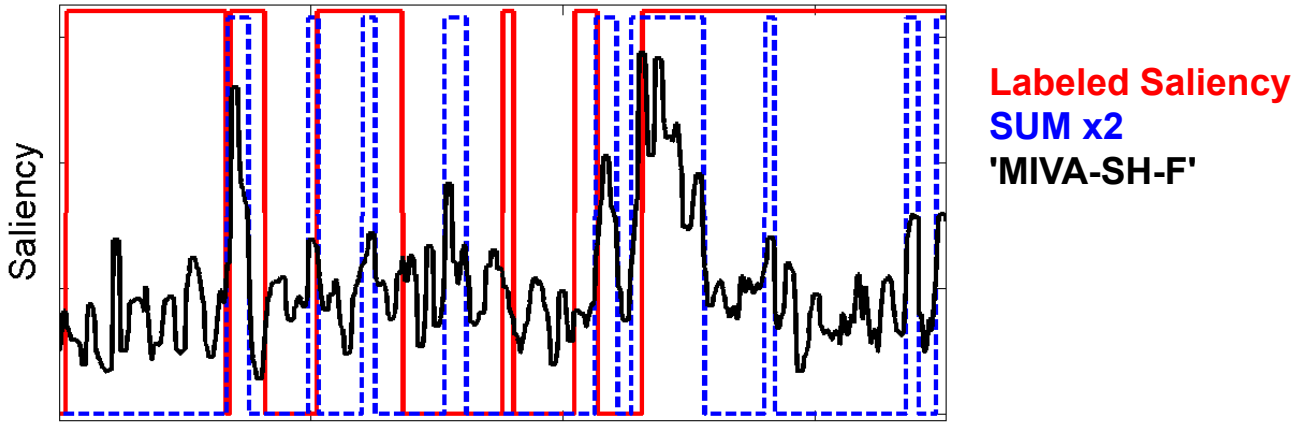
G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos and Y. Avrithis, *Video Event Detection and Summarization Using Audio, Visual and Text Saliency,* ICASSP 2009.

COGNIMUSE Database

- **7** Academy awarded movies

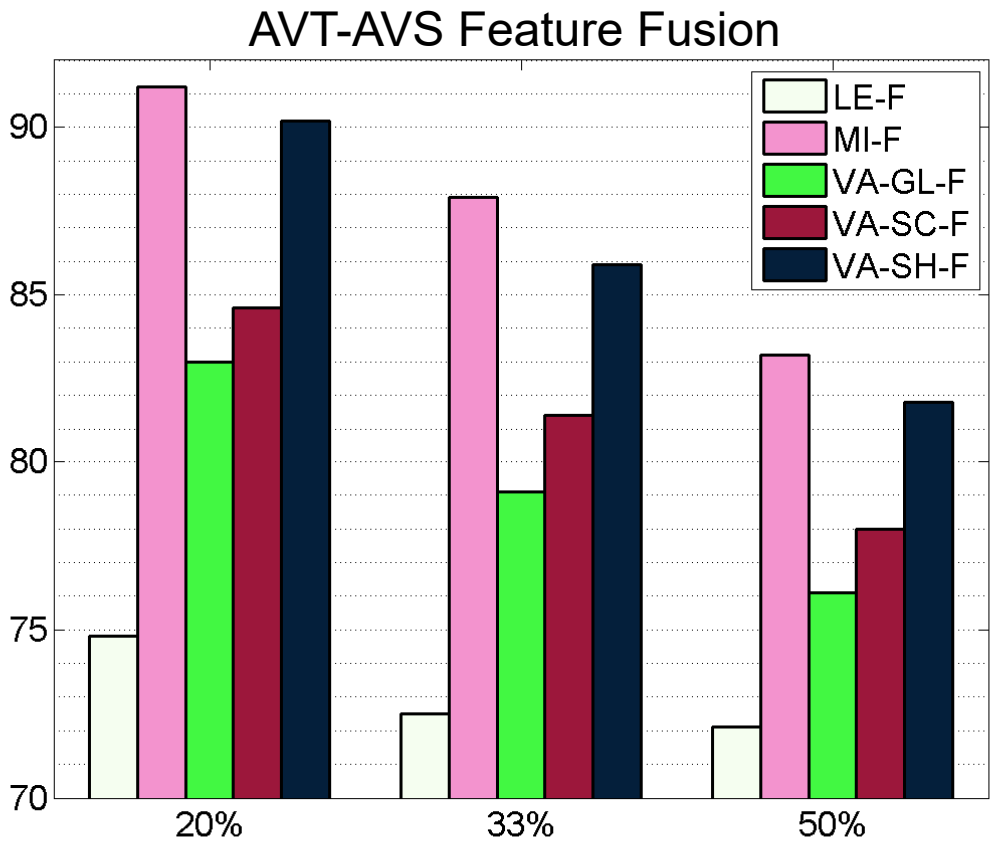
  ca. 30 min. duration segments (on average 13 scenes/movie, 560 shots/movie)
  - **GLA**, *"Gladiator"*: DreamWorks SKG, 2000
  - **BMI**, "*A Beautiful Mind":* Imagine Entertainment, 2001
  - **CHI,** *"Chicago"*: Miramax Films, 2002
  - **LOR**, *"Lord Of the Rings III: The Return of the King"*: New Line Cinema, 2003
  - **CRA,** *"Crash"*: Bob Yari Productions, 2005
  - **DEP,** *"Departed"*: Warner Bros. Pictures, 2006
  - **FNE**, *"Finding Nemo"*: Walt Disney Pictures, 2003

- Skimming rates: c = 20%, 33%, 50% (x5, x3, x2 real time summaries)
- Correspondence with manually labelled saliency



Objective Evaluation
Automatic Movie Summarization (COGNIMUSE Database)

Best Four Fusion schemes with GL-N + baseline (LE-F)
(in terms of frame-level precision)



[G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, *Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention,* IEEE Trans.-MM, 2013]

Demo: Movie Summaries (Bottom-up, Fusion: System #1)

LOR VA-SH-F, rate: x5 (6:50 min from 37:33 min)
Inform: 78.7 %
Enjoy: 80.9 %

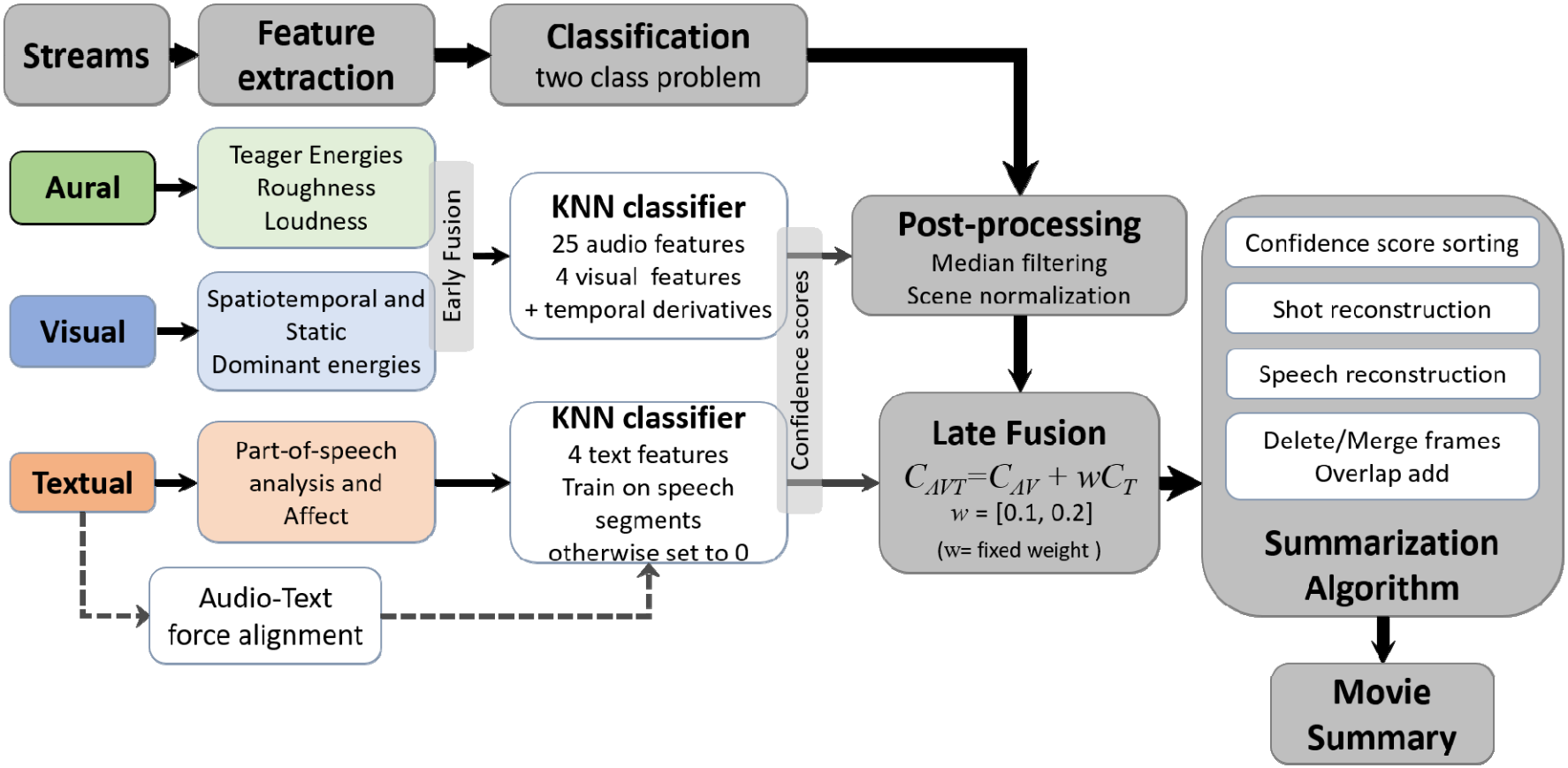
FNE MI-F, rate: x5 (5:07 min from 30:17 min)
Inform: 74.1 %
Enjoy: 78.3 %

[G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, "*Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention*" IEEE Trans.-MM, 2013]

Movie Summarization System #2 (Improved Frontends, Learning)

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.

Summarization System Overview (Learning)



[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]

Audio Analysis

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

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

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

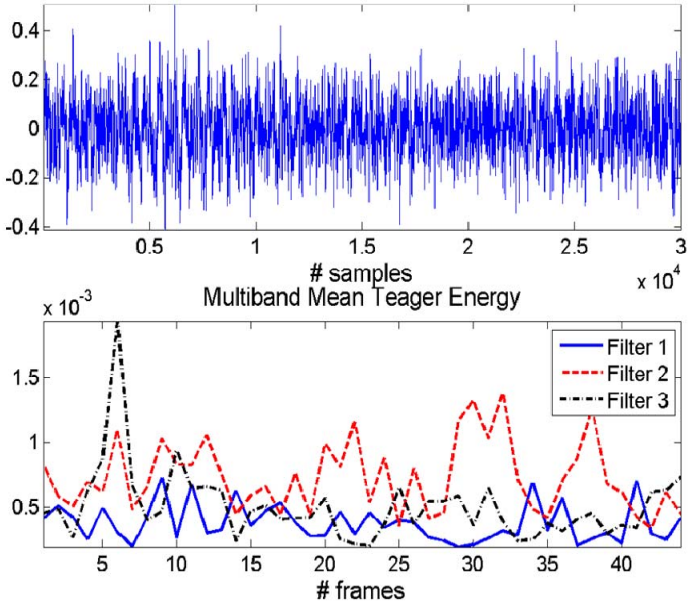
$$\frac{\Psi\left[x(t)\right]}{\sqrt{\Psi\left[\dot{x}(t)\right]}} \simeq |\alpha(t)| \qquad \frac{\sqrt{\Psi\left[\dot{x}(t)\right]}}{\sqrt{\Psi\left[x(t)\right]}} \simeq \omega(t)$$

narrow-band→ Filterbank of 25 Mel arranged Gabor filters

AM-FM model

Auditory Features

- 25 Gabor Energies
- Roughness
- Loudness



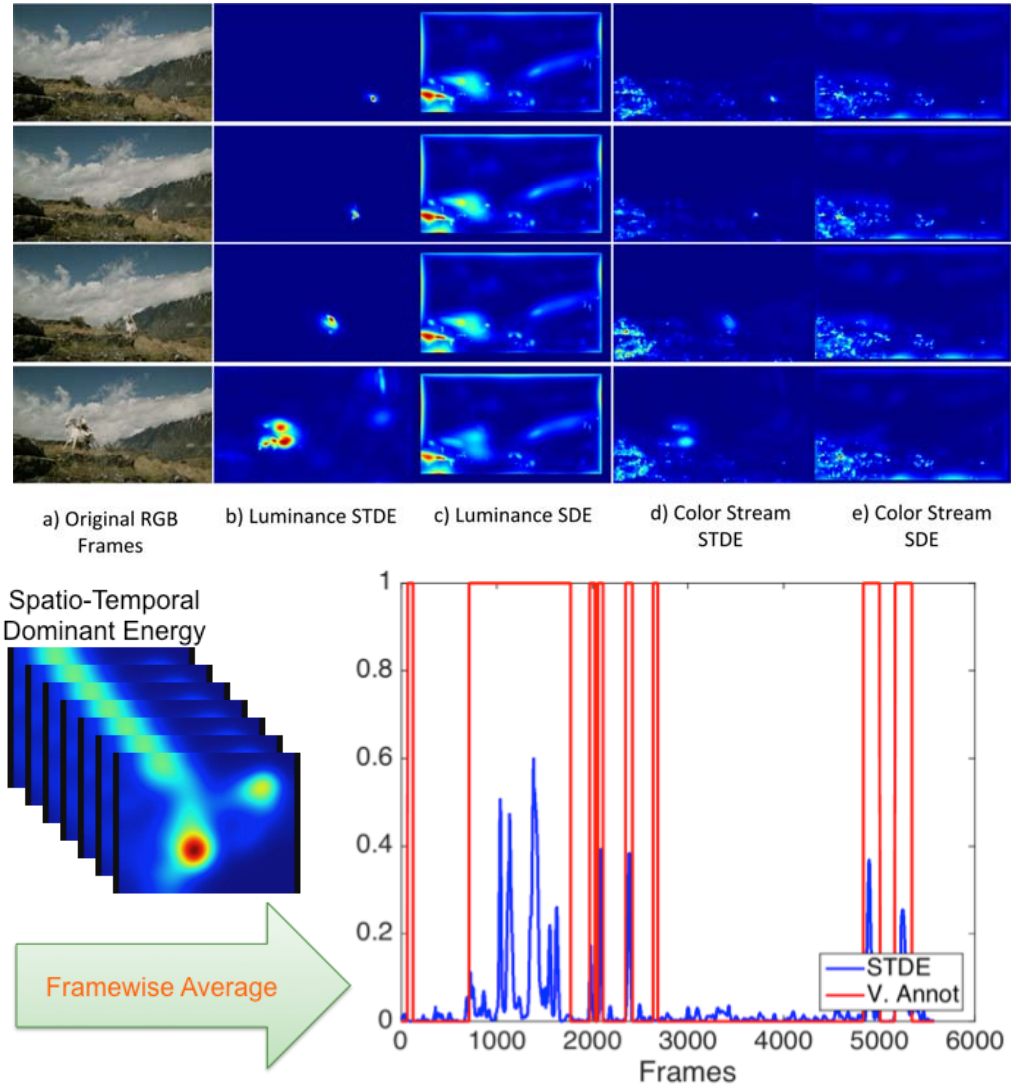
[A. Zlatintsi, E.Iosif, P. Maragos and A. Potamianos. *Audio Salient Event Detection and Summarization using Audio and Text Modalities*, EUSIPCO, 2015]

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
  - 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]

fMRI Validation of Audio-Visual Saliency Model

- Perceptual plausibility of computational models for visual & auditory saliency
- Brain activation data during stimulation
- fMRI data from complex video stimuli

[G. Panagiotaropoulou, P. Koutras, A. Katsamanis, P. Maragos, A. Zlatintsi, A. Protopapas, E. Karavasilis and N. Smyrnis, *FMRI-Based Perceptual Validation of a Computation Model for Visual and Auditory Saliency in Videos*, ICIP 2016]

fMRI Experimental Setup

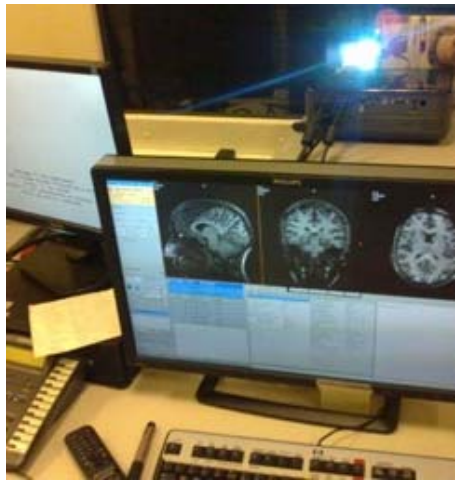
2 experimental setups (free-viewing):

I. “Botswana Lion Brotherhood” documentary [BLB], 15’ – 5 subjects

- Manipulation of image (ON/GRAY/OFF) & sound (ON/OFF)

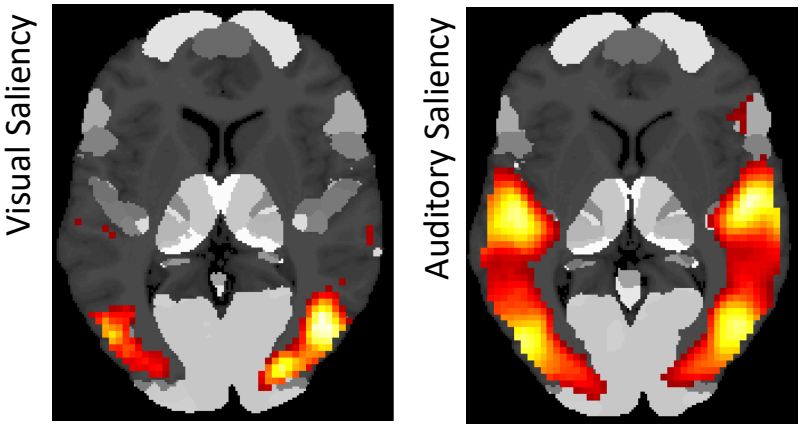
II. “The Departed” film [DEP], 20’ – 6 subjects

- Audiovisual features

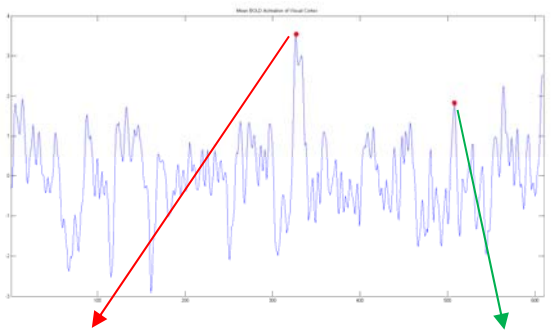


fMRI for Audiovisual Saliency

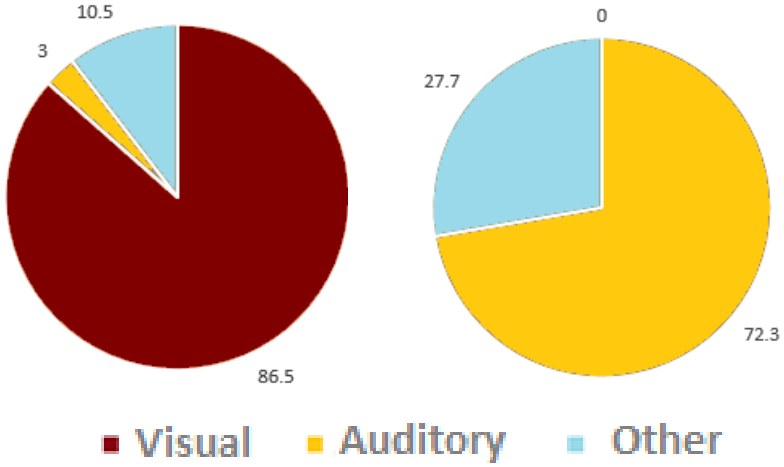
Movie free-viewing [DEP]



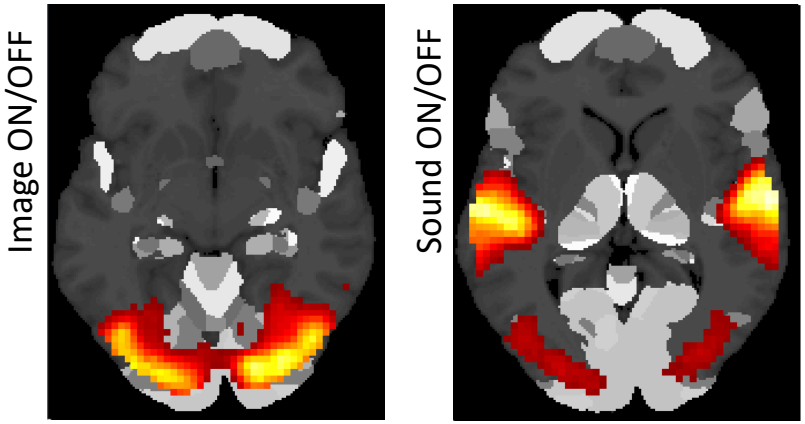
Visually salient action scenes [DEP]



Percentage of active cluster



Documentary ON/OFF design [BLB]



Affective Word-Level Modeling

- High arousal & high absolute valence: good indicators for words related with salient events.
- **Semantic Similarity $Q^H(w_i,w_j)$:**
  - Semantically similar words share similar context.

$$Q^H(w_i,w_j)=\frac{x_i\cdot x_j}{\|x_i\|\|x_j\|},\ x_i=\text{ lexical features}$$

extracted using a window of 2H+1 words centered on every target word wi.

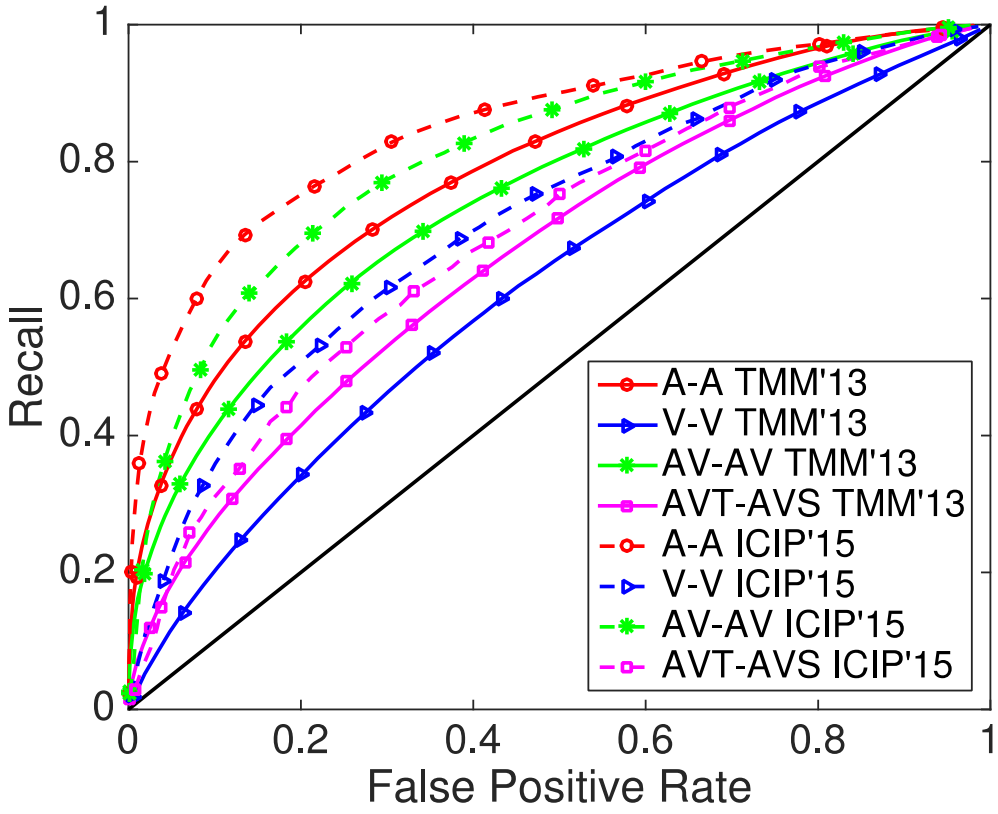
- **Affective Rating of Words w**: valence (v), arousal (a), dominance (d)
  - Linear combination of semantic similarity $S(t_i,w)$ to a set of K seed words and the corresponding affective ratings of seeds (600 entries of the ANEW lexicon).

$$\hat{u}(w)=\lambda_0+\sum_{i=1}^{K}\lambda_i u(t_i)S(t_i,w)$$

  - $t_1 \ldots t_K$ = seed words,
  - u = affective dimension (v, a, d),
  - $u(t_i)$ = affective rating of seed t_i,
  - λi = trainable weight corr. to seed t_i.
  - $S(t_i,w)$ = metric for computation of semantic similarity where $Q^{H=1}$ using text corpus of 116 million sentences.

[N. Malandrakis, A. Potamianos, E. Iosif and S. Narayanan, *Distributional semantic models for affective text analysis*. IEEE TASLP 2013]
[A. Zlatintsi, E. Iosif, P. Maragos and A. Potamianos, *Audio Salient Event Detection And Summarization Using Audio And Text Modalities*, EUSIPCO 2015]

Objective Evaluation: Hollywood Movies

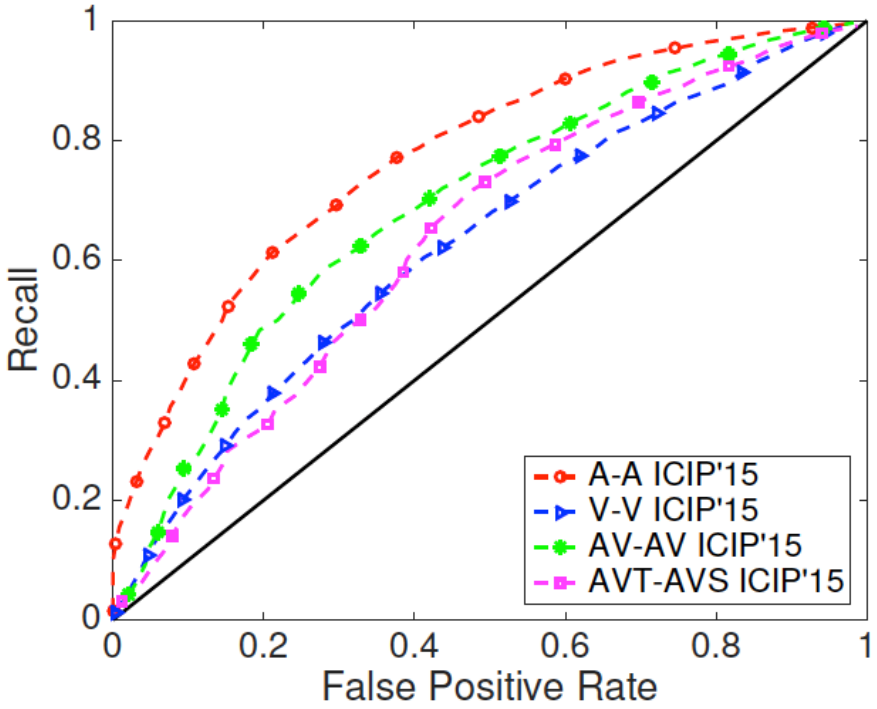


- The proposed system (ICIP 2015) outperforms the baseline MovieSum-1 System (T-MM 2013) for all evaluation setups
- Greater improvement for A-A
- **Improvements** due to:
  - Advanced monomodal frontends, in all modalities
  - Carefully designed movie summarization algorithm that:
    - corrects the boundaries and
    - results in smoother transitions

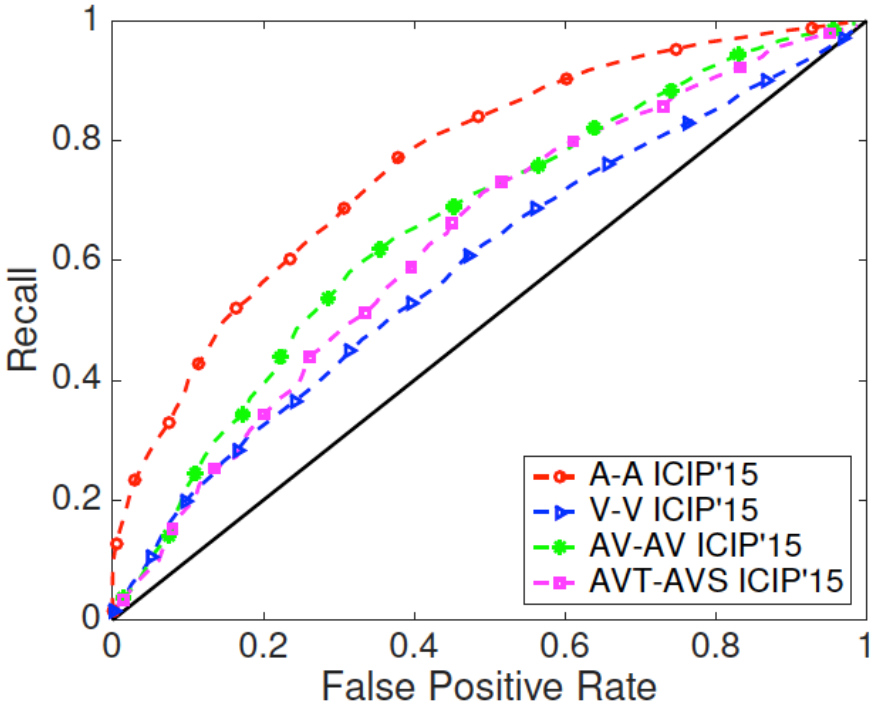
[G. Evangelopoulos et al., *Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention,* IEEE TMM 2013]
[P. Koutras et al., *Predicting Audio-visual Salient Events based on A-V-T Modalities For Movie Summarization,* ICIP 2015]

Objective Evaluation: GWTW Movie

Trained on 7 movies

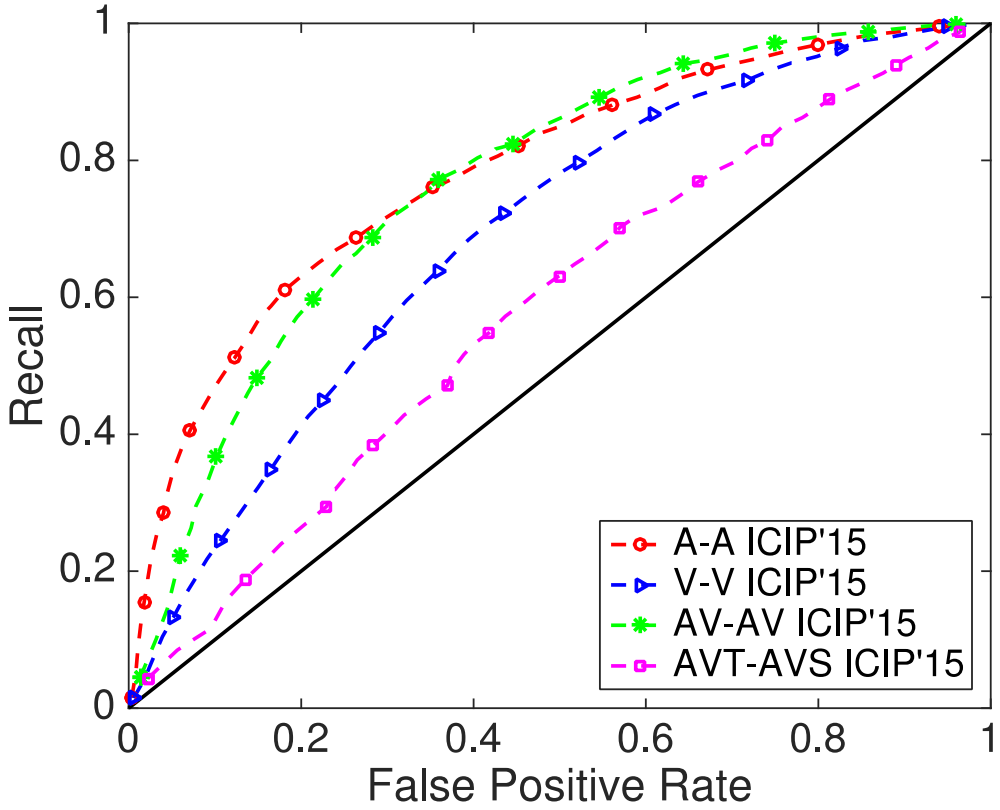


Trained on 7 movies & 5 travel doc



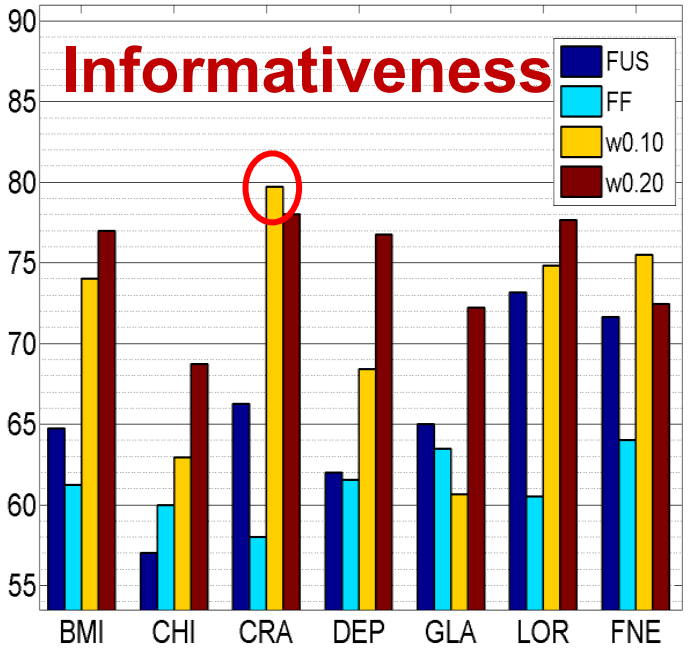
- Best performance for A-A evaluation
- Text seems not to improve the AV case
- Better performance when the training is performed on movies only (since the specific travel doc are unstructured)

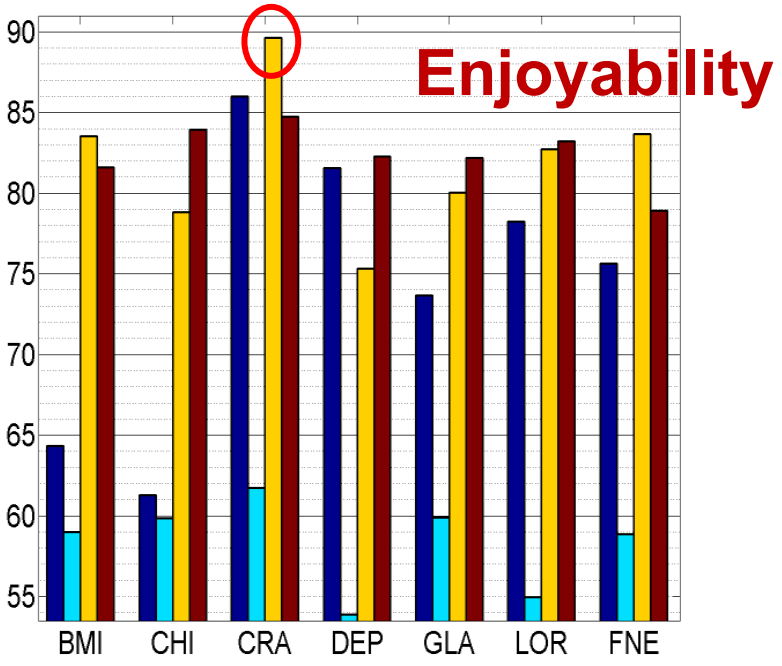
Objective Evaluation: Travel documentaries



- Best performance for A-A evaluation on shorter summaries
- AV fusion slightly better for longer summaries
- Text seems to worsen the AV case significantly (ca to random), probably due to:
  - the unstructured dialogues,
  - every day and slang language
- Keep in mind: that different things are important in “travel” doc. compared to movies

Subjective Qualitative Evaluation





- Setup: Summaries x5, ca. 6 min, 20 users
- Evaluation on:
  - **T_W0.1**: text weight T_W = 0.1
  - **T_W0.2**: text weight T_W = 0.2
  - **FUS**: baseline fusion method
  - **FF**: fast-forward (sub-sampling 2 sec. every 10 sec.).

Results

- Different T_W is important and related to the movie genre
- Action movies need higher T_W
- Boundary correction contributed to enjoyability:
  - a) smoother transitions and
  - b) semantically coherent events

Part 5: Conclusions

- **COGNIMUSE Database**: Multimodal video database with multi-facet annotation, incl. saliency, semantics events and emotion
- Two computational systems for Movie Summarization:

i. Bottom-up multimodal saliency representations of audiovisual streams, integrating signals (audio and visual) and semantics (linguistic/textual)

ii. Improved synergistic approach of audio-visual salient event detection and movie summarization based on a unified energy-based audio-visual framework and a method for text saliency

- Movie summarization system for the production of automatic summaries, evaluated both objectively and subjectively, accomplishing really good results

Tutorial slides: http://cognimuse.cs.ntua.gr/icassp17
COGNIMUSE dataset: http://cvsp.cs.ntua.gr/datasets

Video Summaries: Hollywood Movies

CRA (w0.1) ca 20%, ca 5'30" informativeness up to 80%

CHI (w0.2) ca 20%, ca 7' enjoyability up to 85%

Video Summaries: Travel Docum. & GWTW

AR London ca 16% ca 3'40"

GWTW ca 3% ca 3'
(3min from full duration movie)