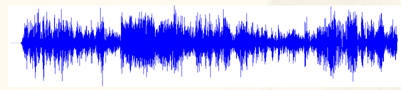
Summarization and Emotion Tracking in Movies

Alexandros Potamianos, National Tech. Univ. of Athens

Outline

* What is a movie summary?
* Attention and Saliency
* Audio, Visual and Text Saliency
* A bottom-up approach to summarization
* Examples
* Credits:
  - TUC team: A. Potamianos, N. Malandrakis, G. Skoumas
  - NTUA team: P. Maragos, G. Evangelopoulos, N. Zlatintsi, K. Rapantzikos

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)

Movie Summarization basics

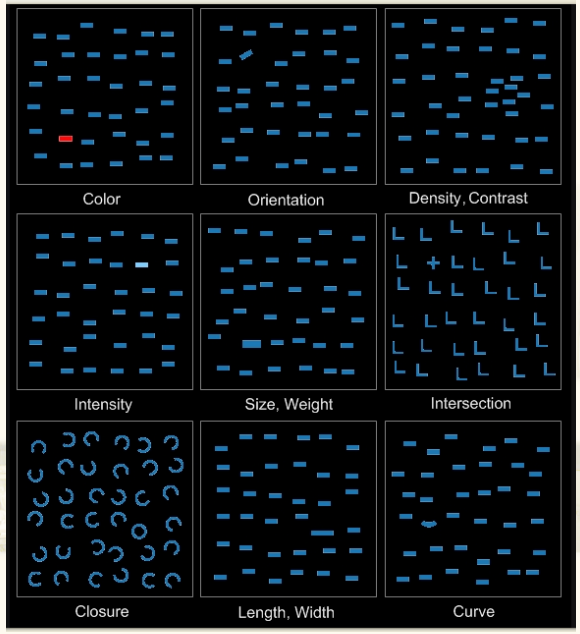
* What is a good movie summary
  - Contains all basic plot elements
  - Contains all important (salient) events
  - Above all: it is itself a good movie
* In essence, a good movie summary should be as informative and enjoyable as the original movie
* Top-down (semantic) and bottom-up (cognitive/perceptual) approach to movie summarization

Example: x2 compression

Cognition and Attention

* What grabs our attention?
  - Salient events
* Attention and Perception:
  - A **simple** perceptual algorithm
  - Quickly identify relevant (to survival) information
* Features extracted via low level signal processing
* The attention/saliency relationship is used in multimedia production

What Grabs Your Attention in an Image?



from http://www.feng-gui.com

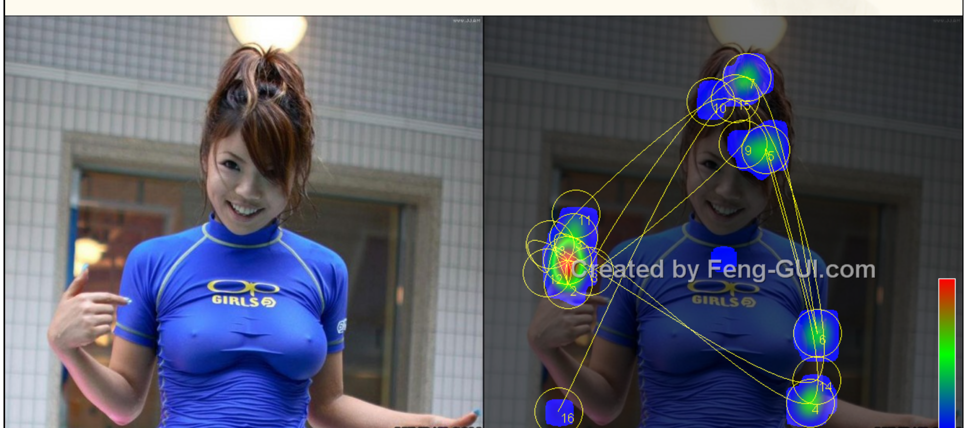
More on Attention and Saliency

* Similarly for video: low level attention features are motion (direction, velocity), flicker
* Such low level features capture about 75% of "events" in images
* How do we capture the rest?
  - Multimodality (redundancy)
  - Semantics of multimedia

Attention Models: Good Example

example from http://www.feng-gui.com

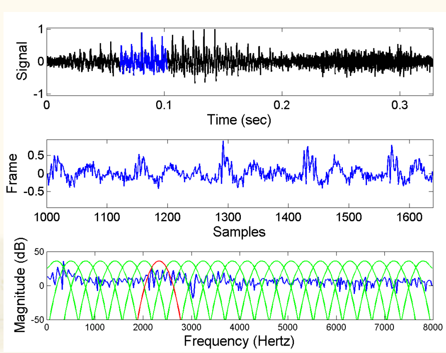
Attention Models: Bad Example

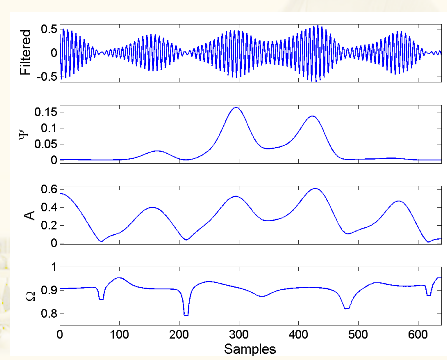


Attention Models and Saliency

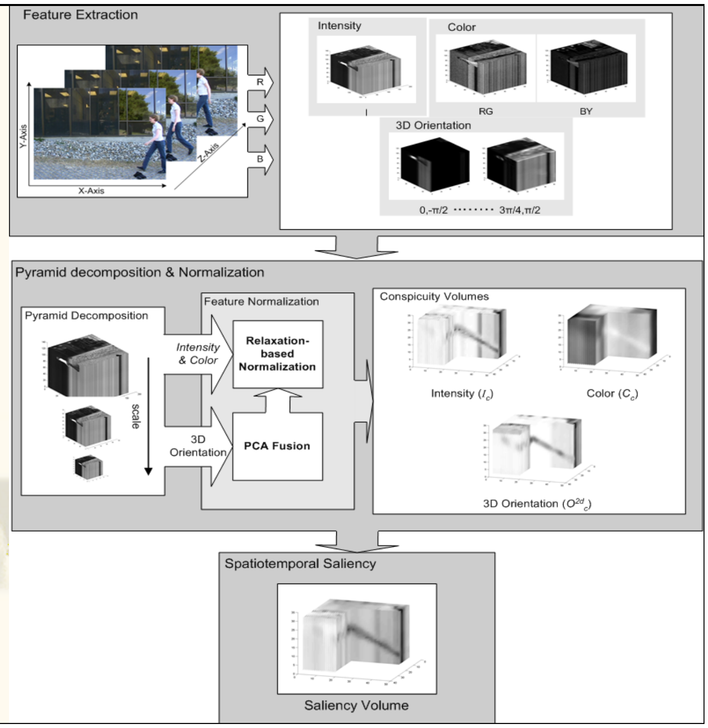
* Attention model of video streams
* Saliency measures:
  - Aural: energy of multi-frequency band features
  - Visual: multi-scale intensity, color and motion
  - Text: part of speech assignments
* Fusion on a single audio-visual-text saliency metric

Audio Saliency Features

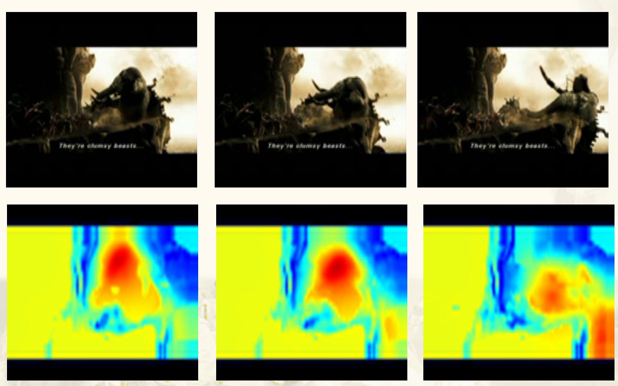




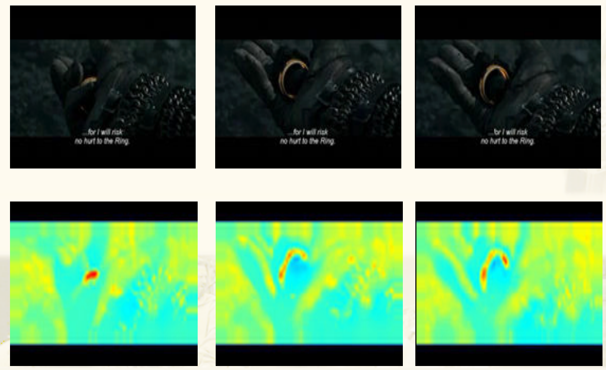
Visual Saliency Features



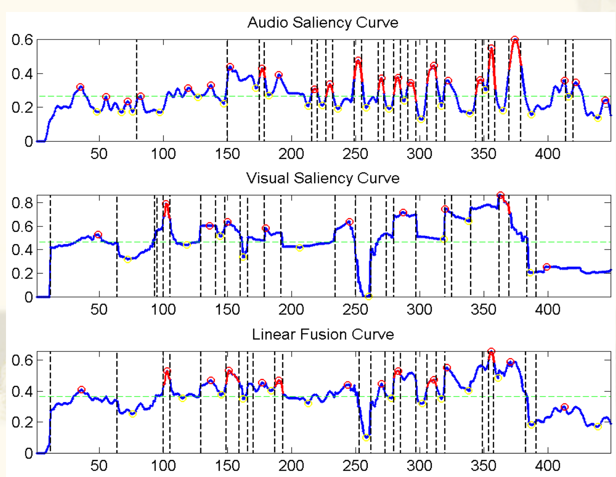
Visual Saliency Example (1)



Visual Saliency Example (2)



Linear AV Saliency Fusion



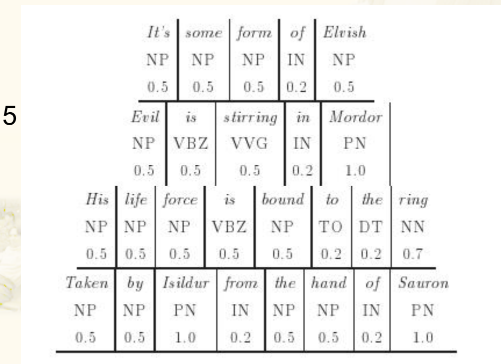
Salient Frames (AV Peaks)

Text Saliency Computation

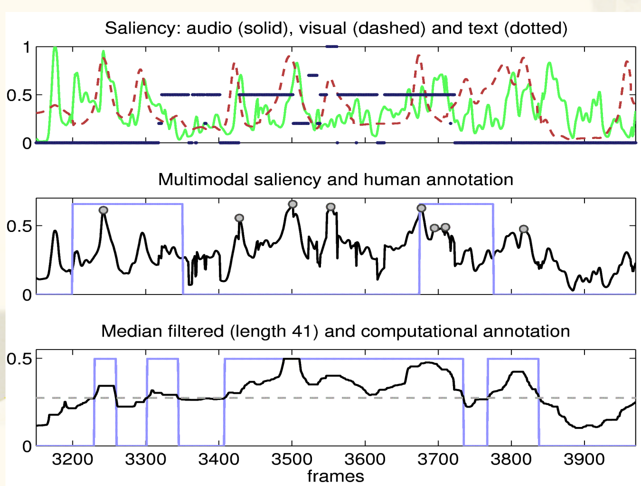
- Extract movie transcript from subtitle file
- Perform part of speech tagging.
- Align text and audio using a speech recognition system
- Assign text saliency value to each frame based on the parser POS tags
- Combine audio, visual and text saliency scores
  - AVT saliency computed at the frame level

POS Saliency Scores

* We consider 6 subclasses and we weight them according
  - Proper Nouns → 1
  - Nouns → 0.7
  - Noun Phrases → 0.5
  - Verbs → 0.5
  - Adjectives → 0.5
  - Stop Words →0.2



AVT Salience via Linear Fusion



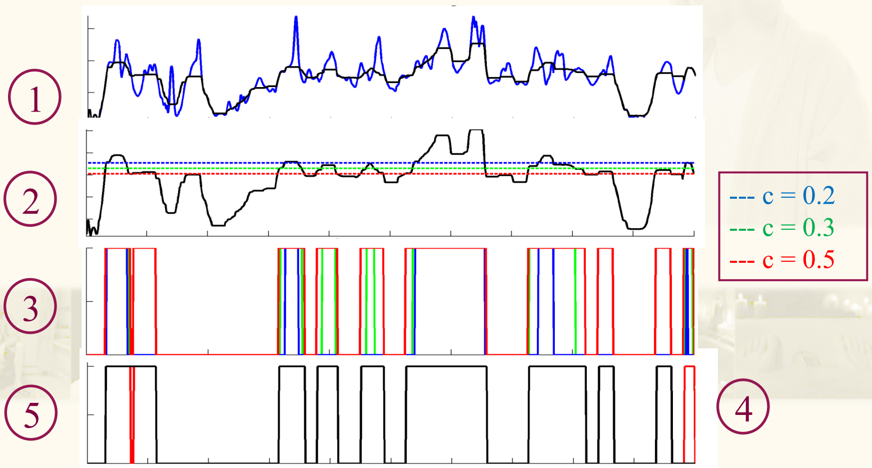
AV Key Frames: 300

Movie Summarization Algorithm

1. Filter: AVSC with median of length $2M + 1$.
2. Threshold choice
3. Selection: segments
4. Reject: segments shorter than N frames
5. Join: segments less than K frames apart
6. Render: Linear overlap-add on L video frames and audio

Evaluation: $M = N = 20$, $K = L = 10$ (videos at 25 fps).

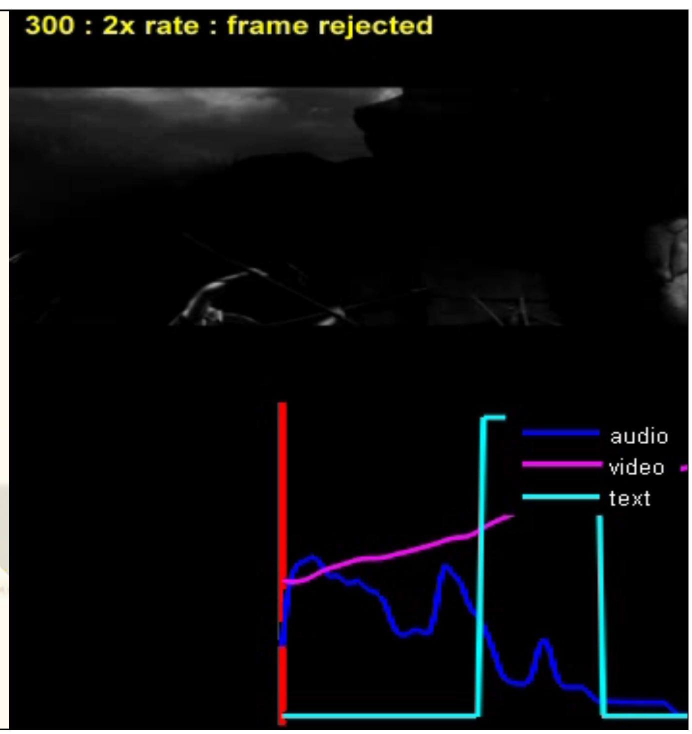
Movie Summarization Algorithm (2)



Summary annotated with AVT Saliency

Grey – Rejected

Color- Accepted in summary



Subjective Evaluation

- 3 clips with duration from 5 to 7 min from the "Lord of the Rings I" (LOTR1), "300" and "Cold Mountain" (CM).
- Skims for $c = 0.5, 0.3, 0.2$ (x2, x3, x5 real time)
- 11 naive users rated originals & skims w.r.t. included information and aesthetics
- 0-100 scale for informativeness and enjoyability

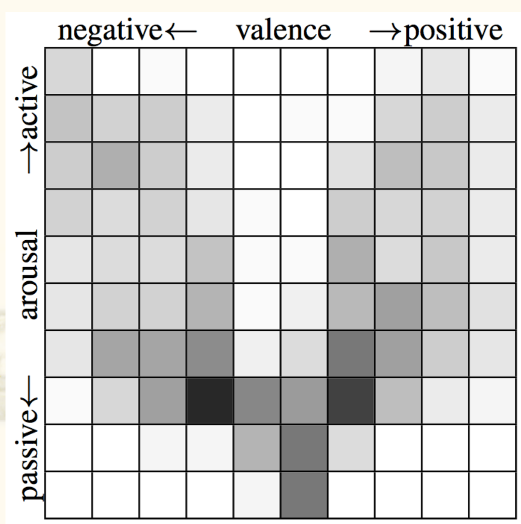
Evaluation: AV vs AVT Summaries

	Multimodal (AVT)			**Relative change (AV)**		
Video	x2	x3	x5	x2	x3	x5
			Informativeness			
LOTR1	86.2	76.3	60.7			
300	86.8	77.9	61.4			
CM	78.4	67.1	59.3			
			Enjoyability			
LOTR1	89.0	80.8	71.1			
300	92.4	86.0	68.6			
CM	84.5	76.8	71.8			

Emotion Tracking

* A continuous 2-D space of emotions
  - Arousal: strength of emotions, related to attention
  - Valence: positive vs. negative emotions
* Affect and audio
* Affect and text
* Visual affect, e.g., emoticons

Affective context of multimedia



Example Frames

Low Activation High

Low Valence High

Emotion Tracking

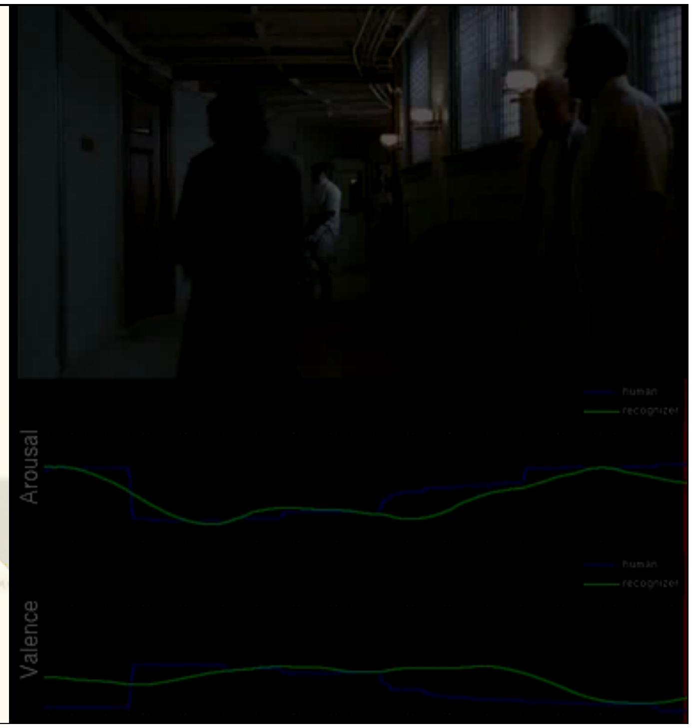
* Quantization of affective space in 7x7 grid
* HMM classifiers built separately for arousal & valence
* Features

Valence	audio video video	12 MFCCs and C0, plus derivatives maximum color value maximum color intensity
Arousal	audio	12 MFCCs and C0, plus derivatives

Affective tracks:
Arousal & Valence

Green– Machine

Blue – Human Annotators (average)

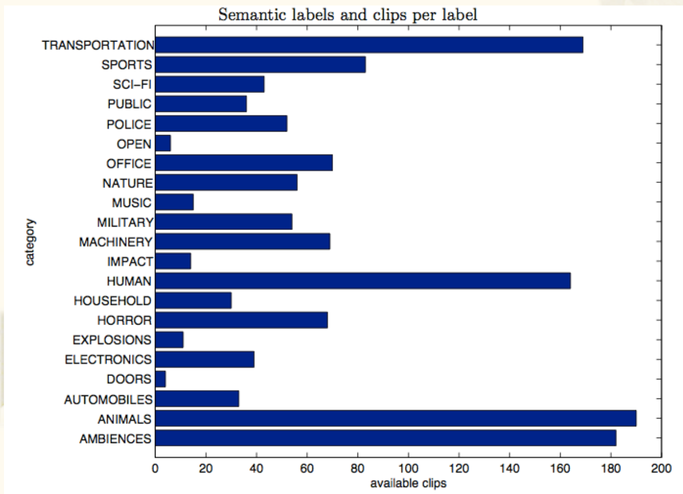


Affective Classification of Generic Audio Clips using Regression Models

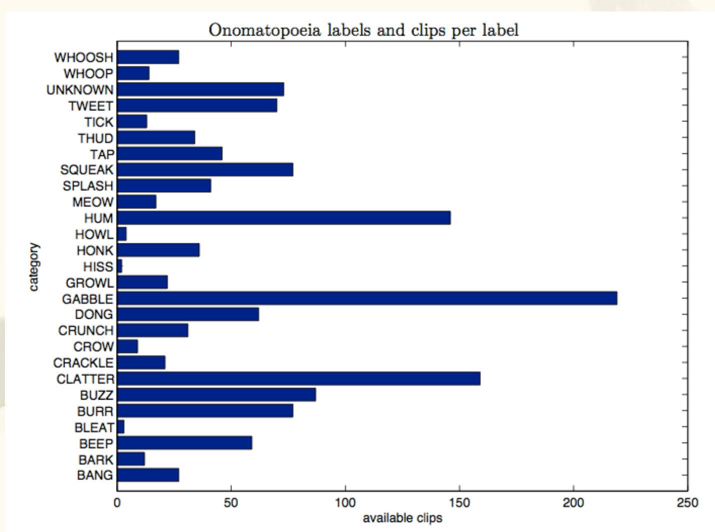
N. Malandrakis, S. Sundaram, A. Potamianos

InterSpeech 2013

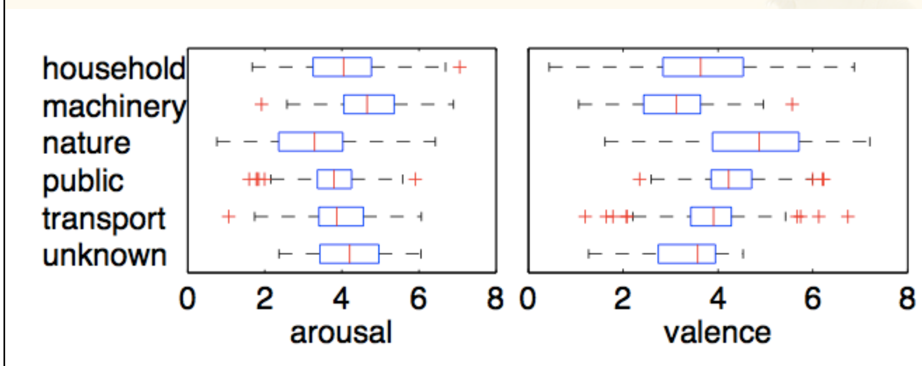
Semantics of Generic Audio I



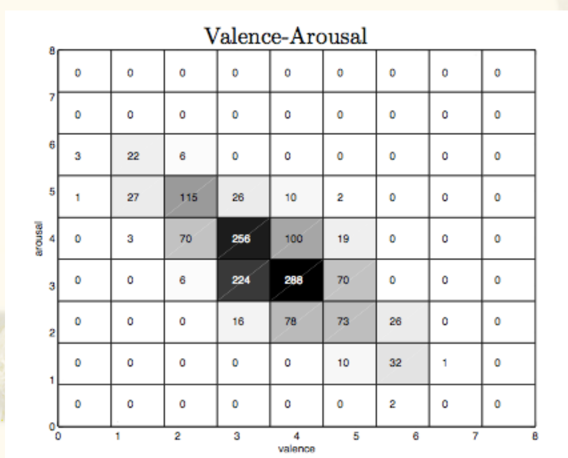
Semantics of Generic Audio II



Overall affective characterization



Distribution of Clip Average Ratings



Inter-annotator agreement

Inter-annotator agreement			
Metric	Arous.	Valen.	Domn.
avg. pairwise correlation	0.52	0.55	0.16
avg. pairwise mean abs. dist.	2.02	1.84	2.32
Krippendorff's alpha (ordinal)	0.39	0.47	0.11
Krippendorff's alpha (interval)	0.39	0.46	0.10
Agreement with the ground truth			
Metric	Arous.	Valen.	Domn.
avg. correlation	0.55	0.60	0.41
avg. mean abs. dist.	1.42	1.18	1.36

Frame level vs Long-Term Features

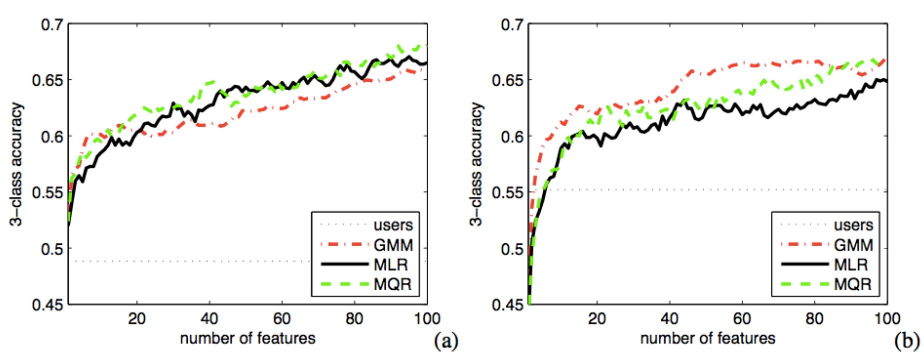
Scope	Low Level. Descr.	Arous.	Valen.	Domn.
frame level	chroma + Δ	0.41	0.45	**0.43**
	log Mel power + Δ	0.44	0.48	0.44
	MFCC + Δ	0.45	0.44	0.43
long term	chroma + Δ	0.41	**0.46**	0.42
	log Mel power + Δ	**0.46**	**0.49**	**0.46**
	MFCC + Δ	**0.48**	**0.48**	**0.45**

Feature Selection

Model	# of features	Arous.	Valen.	Domn.
Users	-	0.55	0.60	0.41
MLR Regression Model	10	0.70	0.67	0.63
	20	0.72	0.70	0.65
	30	0.74	0.71	0.67
	40	0.75	0.72	0.68
	50	**0.75**	**0.73**	**0.69**

3-class Classification Accuracy

Arousal Valence



Conlcusions

* Bottom-up approach to summarization can produce high-quality summaries
  - Redundancy of attention markers in multimodal streams
  - High production value: the attention-saliency loop
* Emotion tracking in movies gives good results with low-level audio (mostly) features