Semantic-Affective Models for Multimedia Data

Alexandros Potamianos

National Technical Univ. of Athens

Univ. of Southern California

Outline

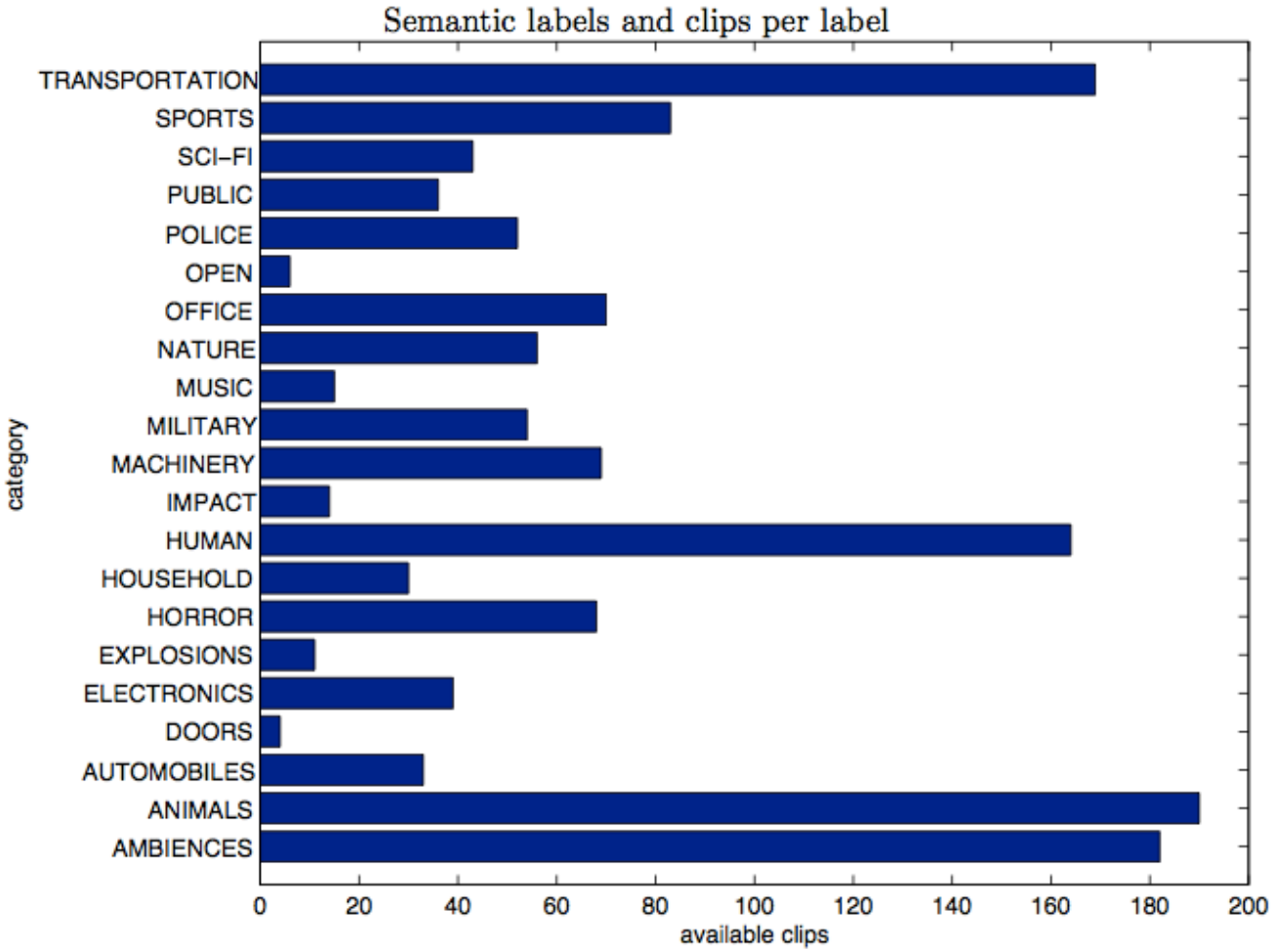
- Affective Modeling
  - Affective Classification of Audio Clips
  - Affective Tracking of Movies
- Multimedia and Cognition
  - Saliency and Attention
  - Representation modeling
- Semantic-Affective Models
  - Symbolic, Associative, Conceptual
  - Representation models in machine learning
  - Our proposal: Audio, Music, Speech
- Grand Challenges

Affective Classification of Generic Audio Clips using Regression Models

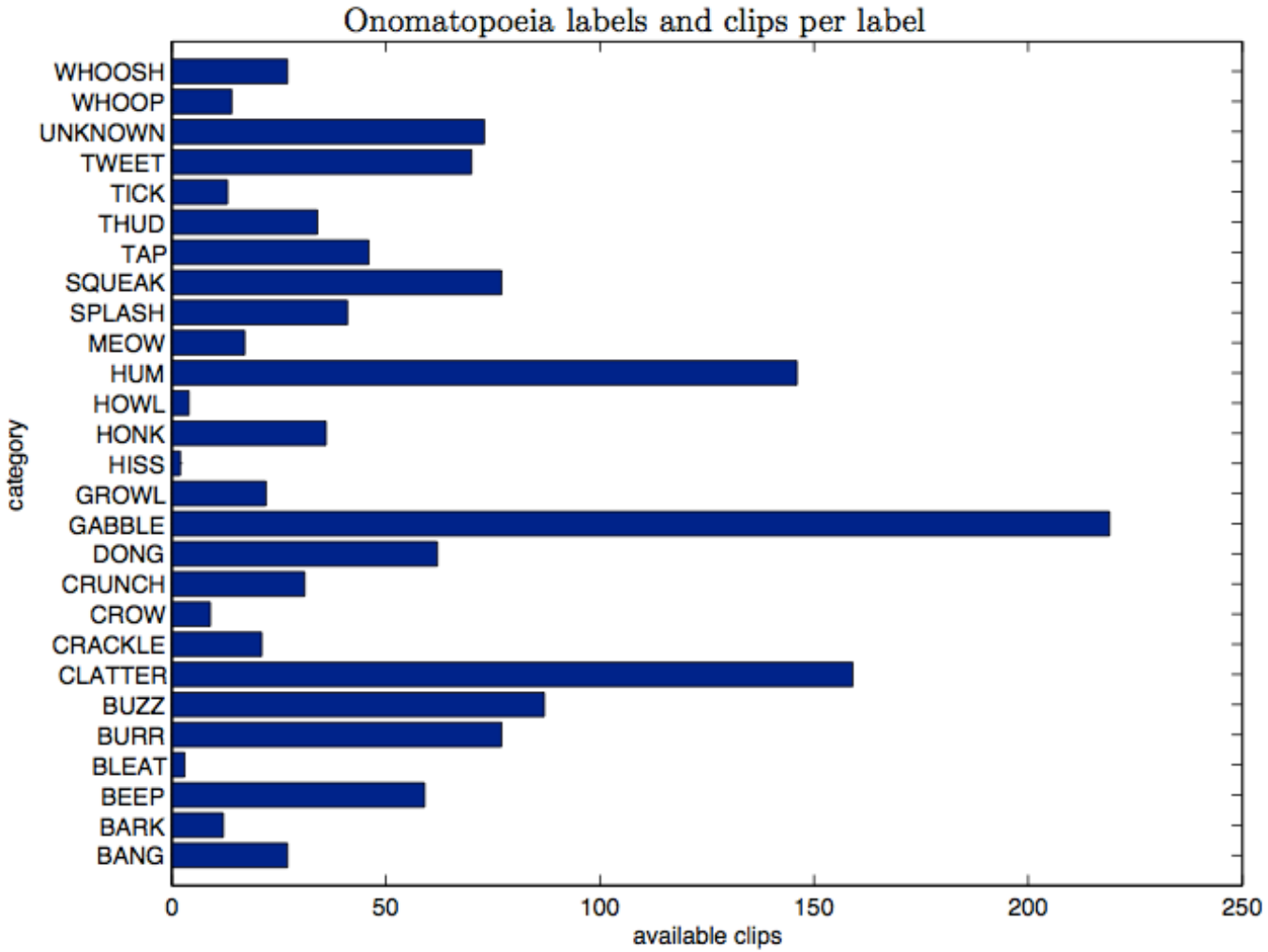
N. Malandrakis, S. Sundaram, A. Potamianos

InterSpeech 2013

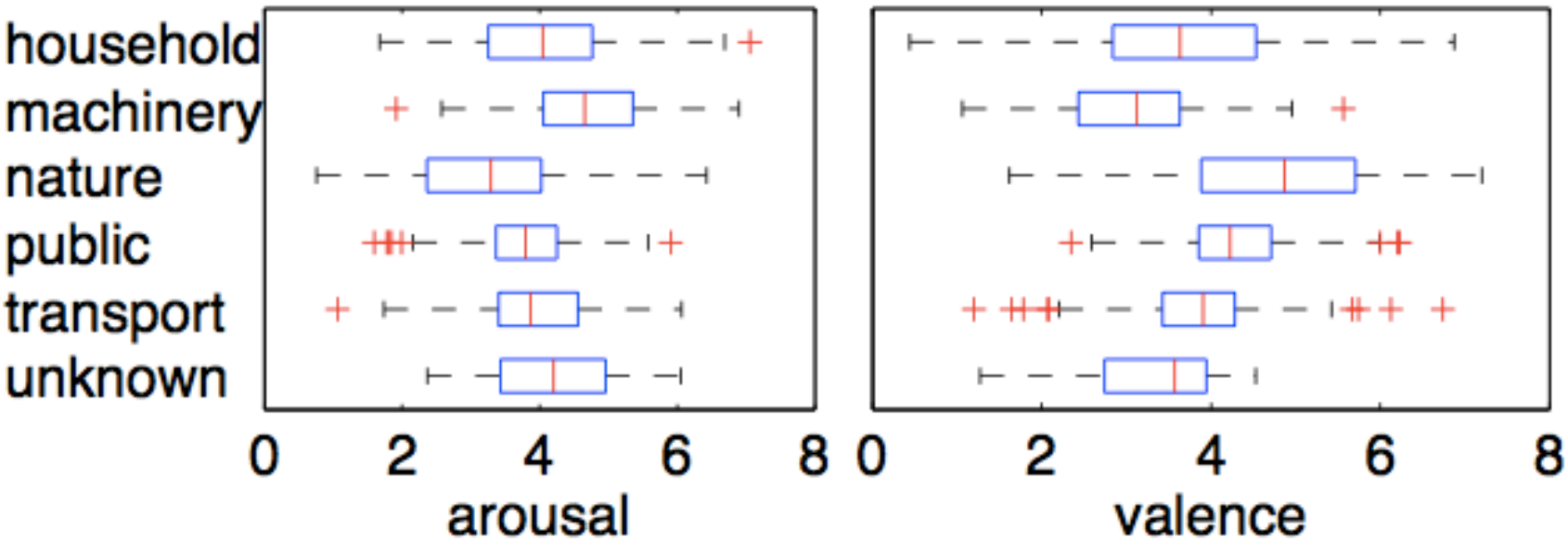
Semantics of Generic Audio I



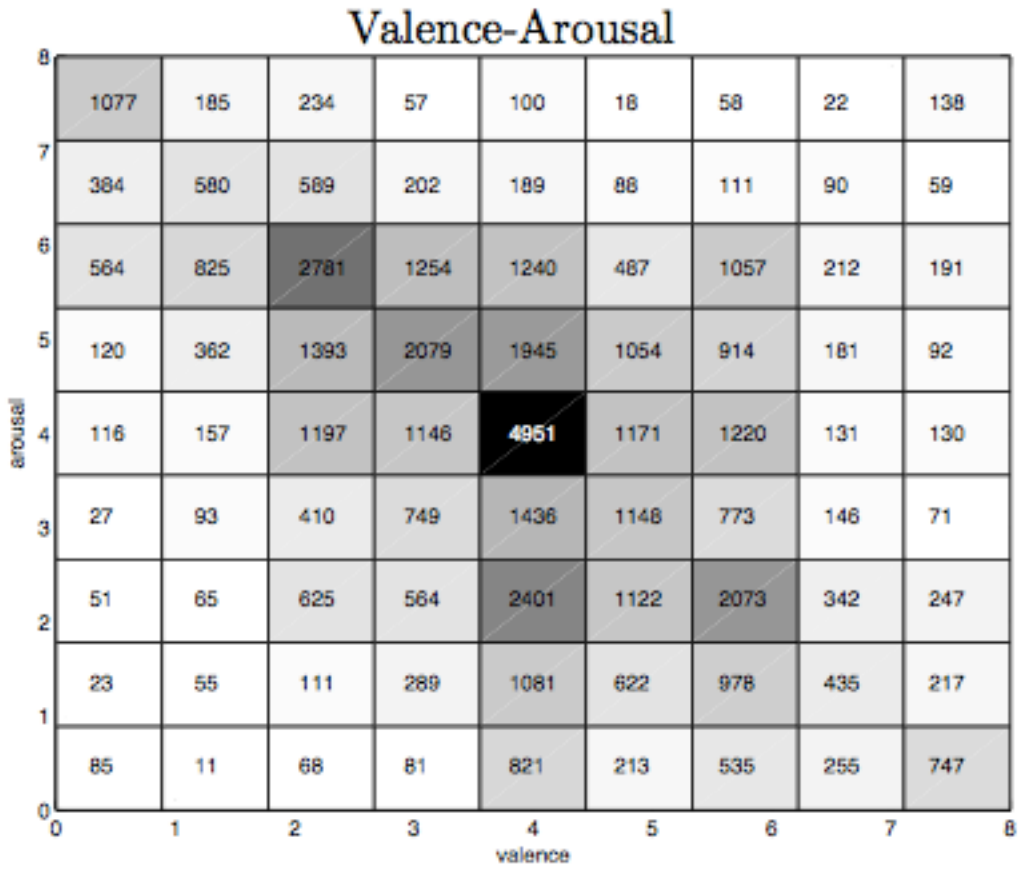
Semantics of Generic Audio II



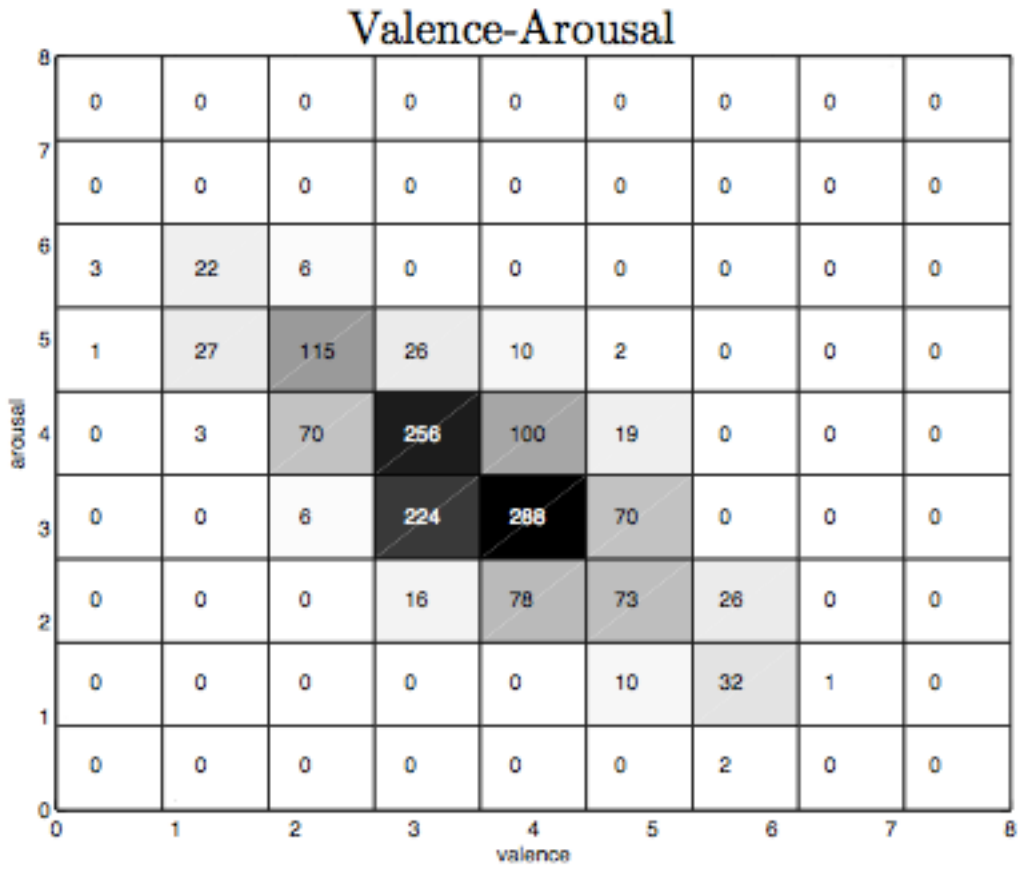
Overall affective characterization



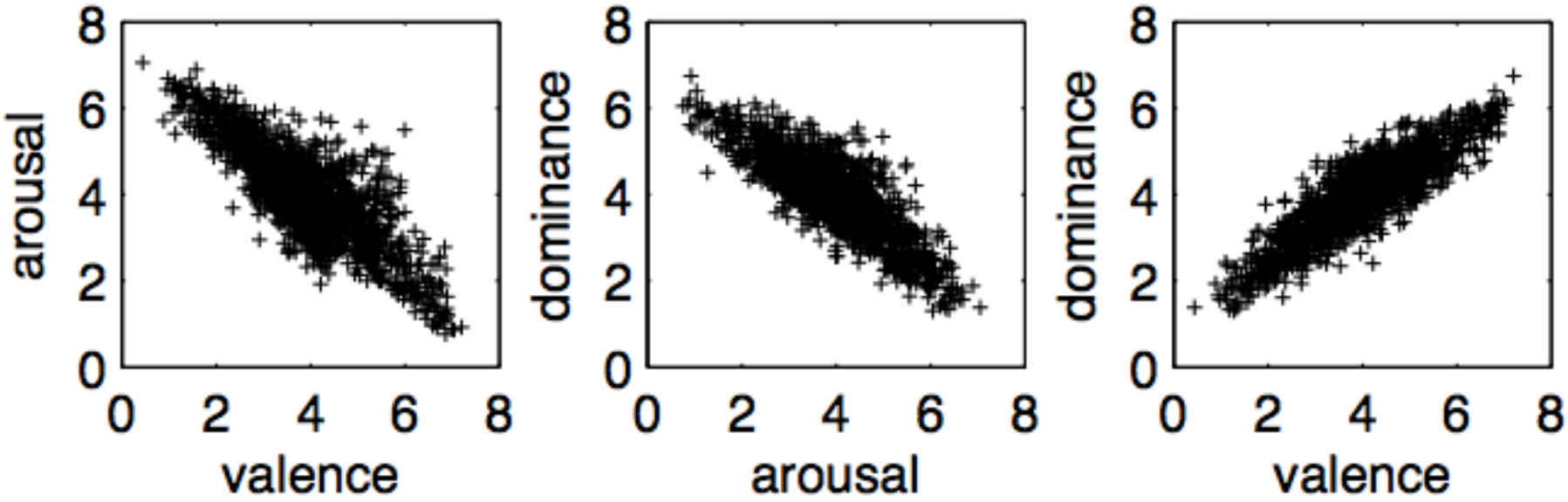
Distribution of All Ratings



Distribution of Clip Average Ratings



3D Affective space correlations



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

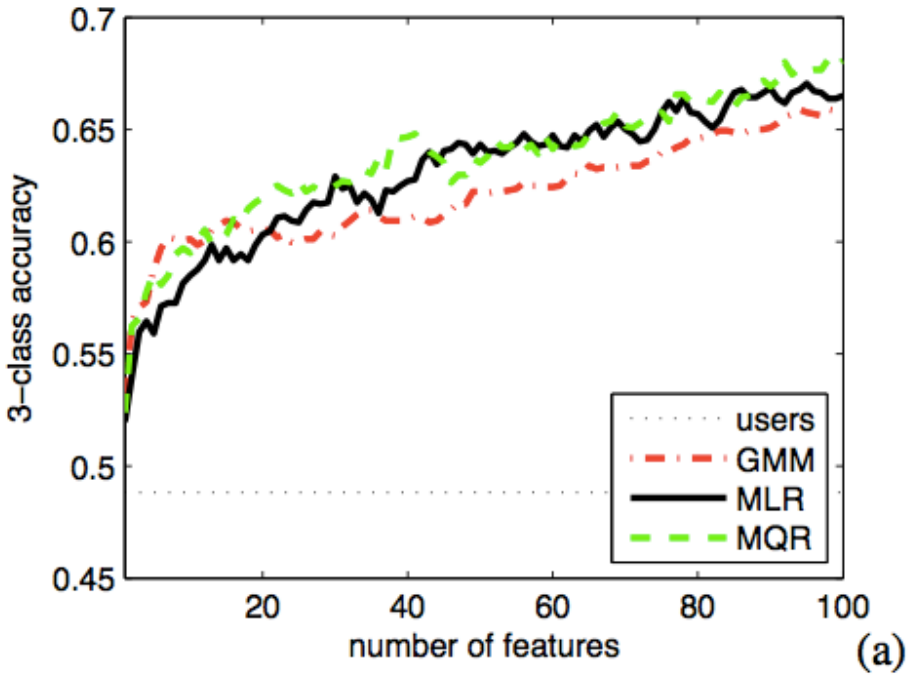
Scope	Low Level. Descr.	Arous.	Valen.	Domn.
frame level	chroma + Δ	0.41	0.45	**0.43**
	log Mel power + Δ	0.44	0.48	0.44
	MFCC + Δ	0.45	0.44	0.43
long term	chroma + Δ	0.41	**0.46**	0.42
	log Mel power + Δ	**0.46**	**0.49**	**0.46**
	MFCC + Δ	**0.48**	**0.48**	**0.45**

Feature Selection

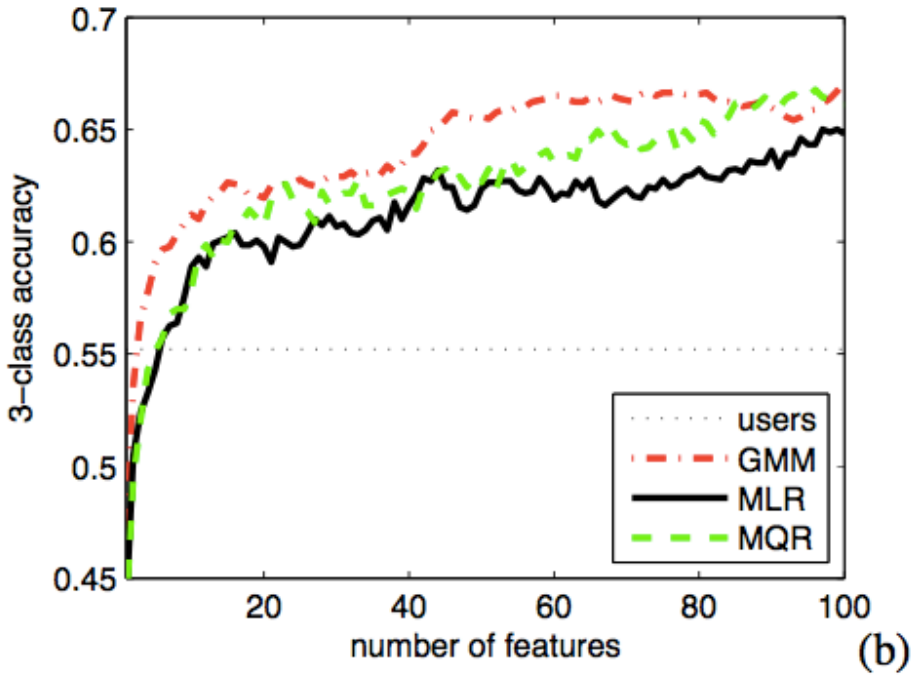
Model	# of features	Arous.	Valen.	Domn.
Users	-	0.55	0.60	0.41
MLR Regression Model	10	0.70	0.67	0.63
	20	0.72	0.70	0.65
	30	0.74	0.71	0.67
	40	0.75	0.72	0.68
	50	**0.75**	**0.73**	**0.69**

3-class Classification Accuracy

Arousal



Valence



A Supervised Approach to Movie Emotion Tracking

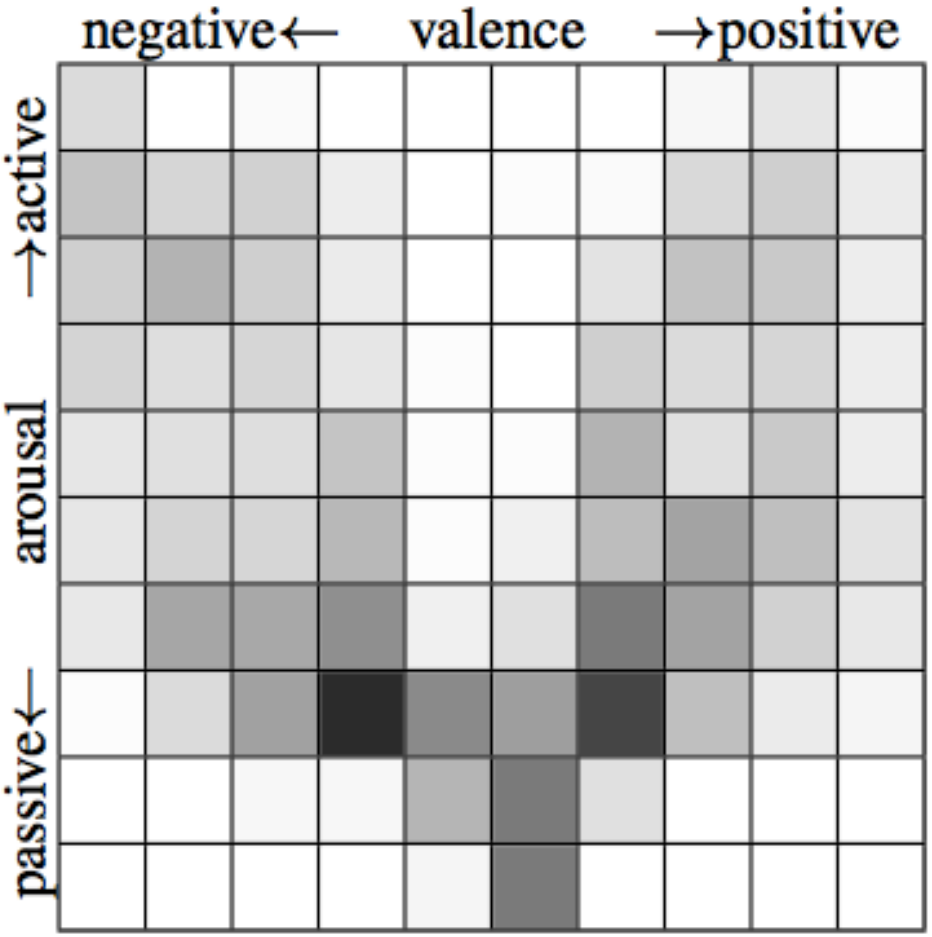
N. Malandrakis, A. Potamianos, G. Evangelopoulos, A. Zlatintsi

ICASSP 2011

Example Frames




Arousal vs Valence Labeled Data



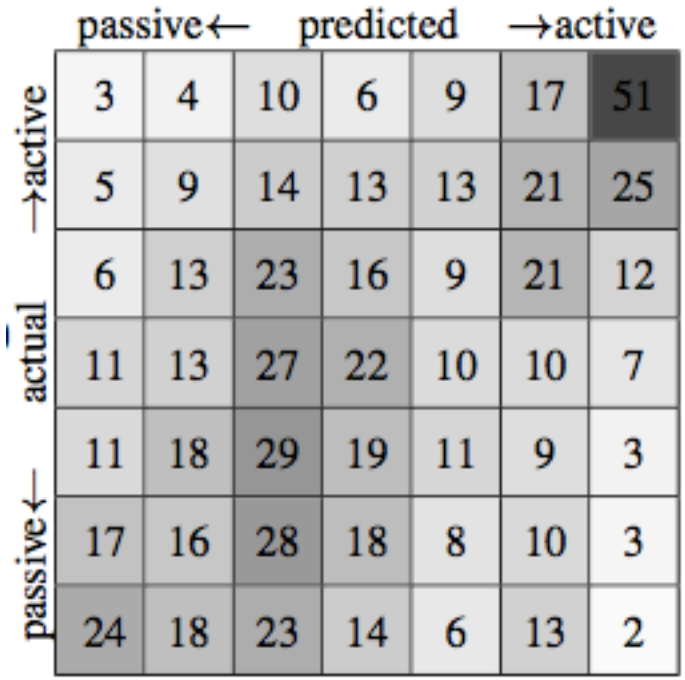
Features and Models

- Continuous-time modeling using HMM models
- Language model used for smoothing
- Features used:

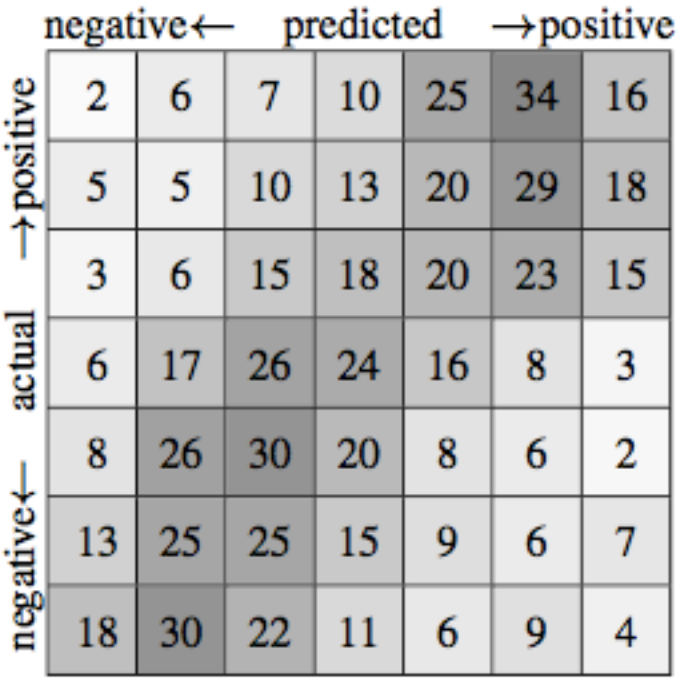
Valence	audio video video	12 MFCCs and C0, plus derivatives maximum color value maximum color intensity
Arousal	audio	12 MFCCs and C0, plus derivatives

Results: Frame Confusion Matrix

Arousal

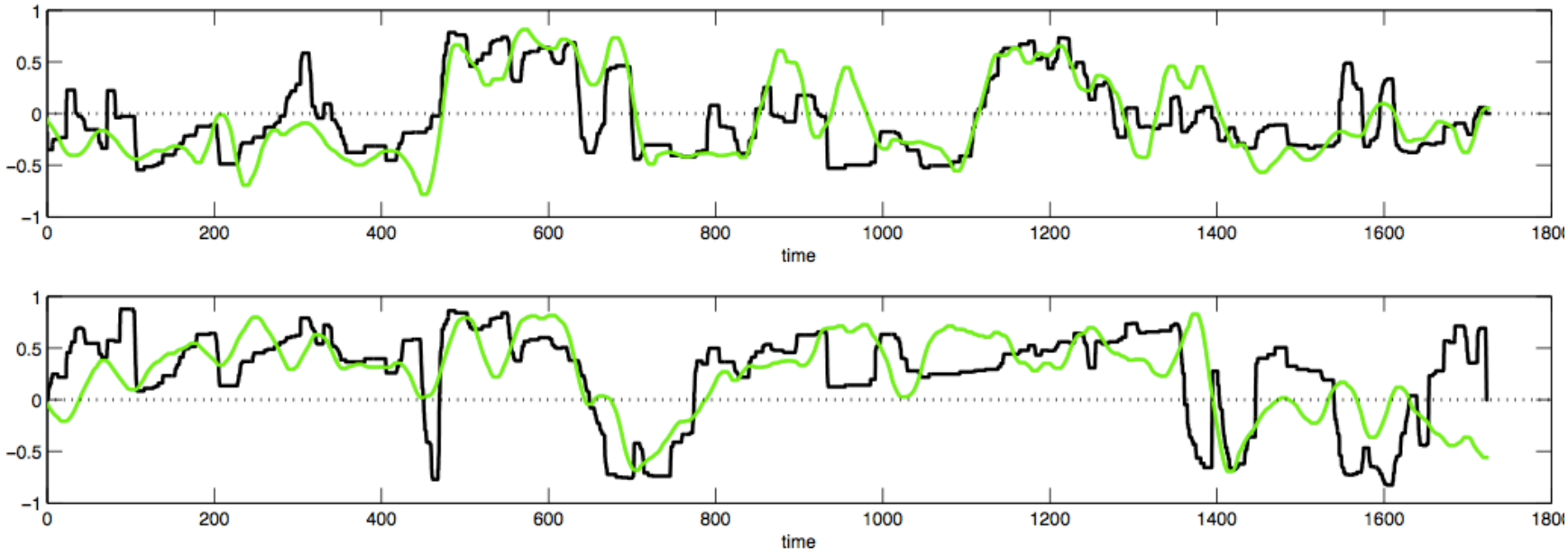


Valence



Continuous-Time Emotion Tracking

Arousal

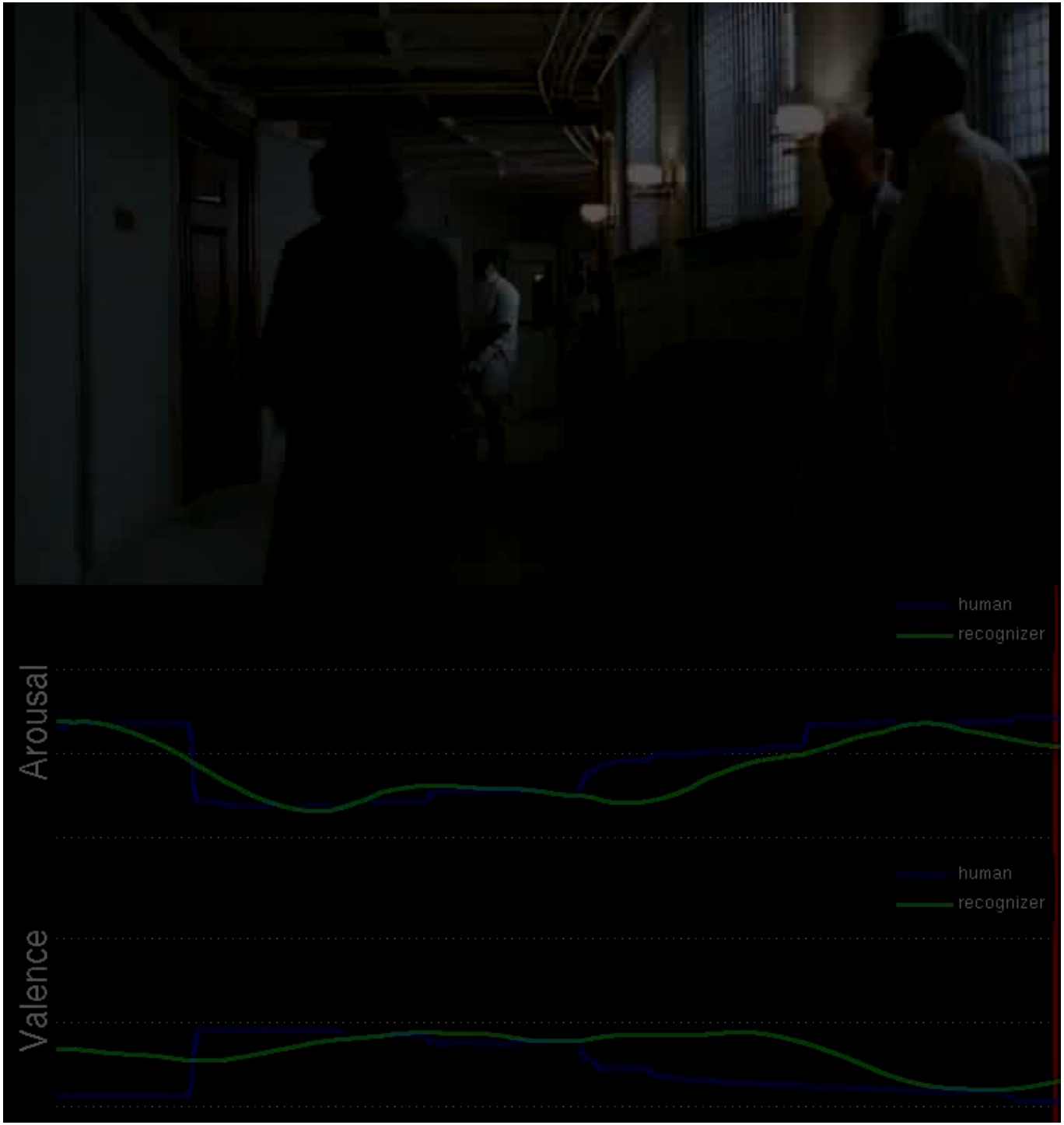


Valence

Affective tracks:
Arousal & Valence

Green– Machine

Blue – Human
Annotators (average)

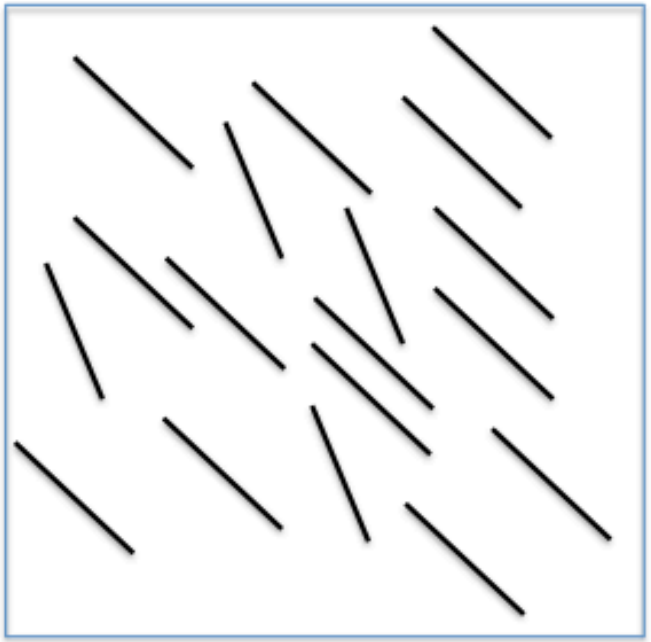


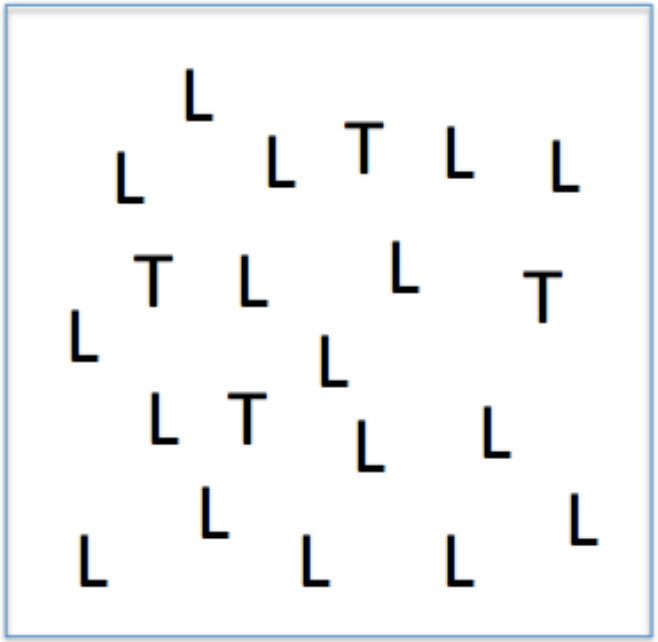
System 1 vs System 2

- Using Kahneman's (and others) formalism:
  - System 1 (intuition): generates
    - impressions, feelings, and inclinations
  - System 2 (reason): turns System 1 input into
    - beliefs, attitudes, and intentions
- Associative relations reside in System 1
- But where do semantic relations reside?

Example

- Example from vision: system 1 vs system 2





Discussion

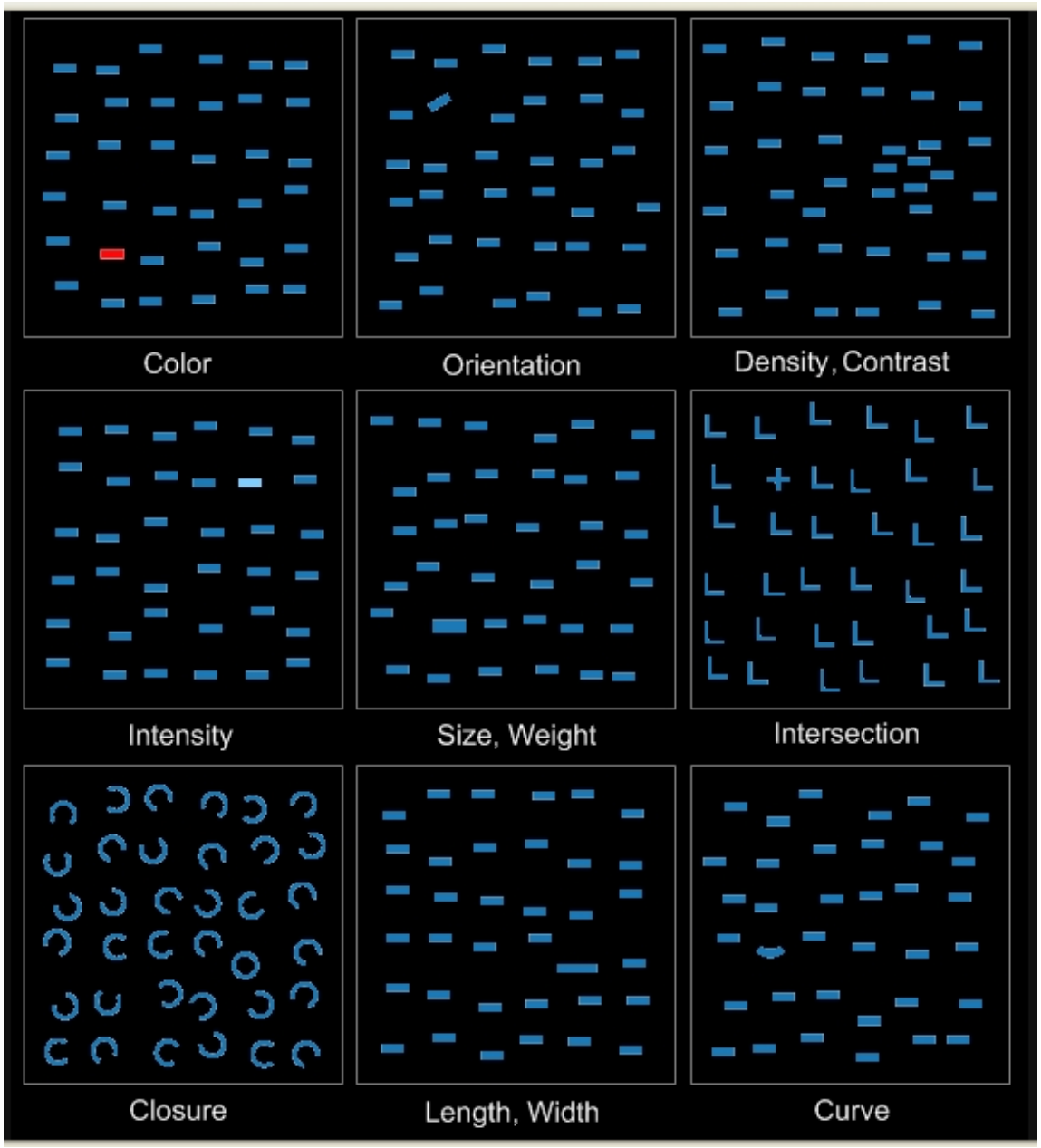
- Affective analysis of generic audio using frame-level features and their statistics
- Affect of movies fusing multimodal cues
- Hard to draw general conclusions about feature selection
- No universal features (except MFCCs!?)
- A detection-based approach for audio processing?

Saliency, Attention and Summarization in Movies

Cognition and Attention

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

What Grabs Your Attention in an Image?



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

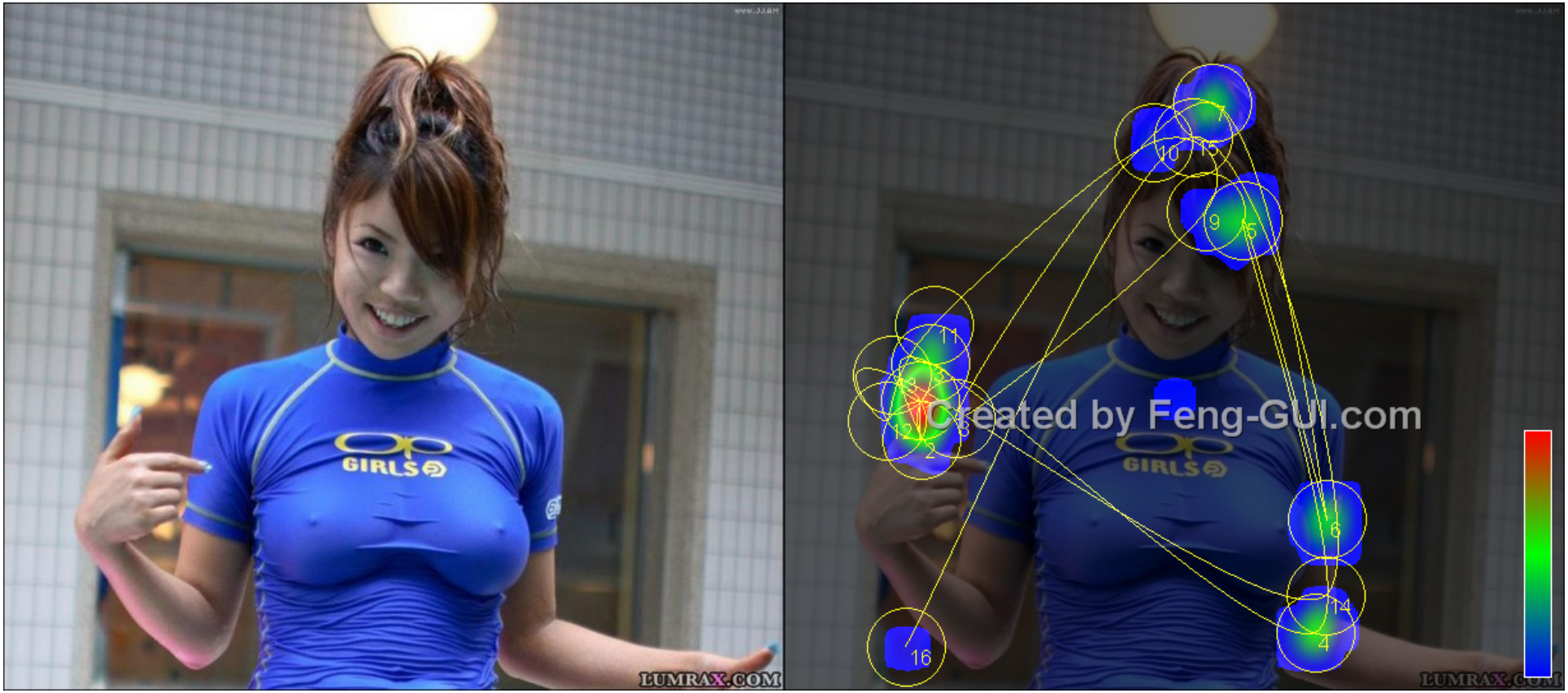
Attention and Saliency

- Audio: rhythm, energy, change of frequency content
- Images over time (video): motion (direction, velocity), flicker
- Such low level features capture about 60-80% of "events" in each modality
- How do we capture the rest?
  - Multimodality (up to 90%)
  - Semantics (top-down selectional attention)

Attention Models: Good Example

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

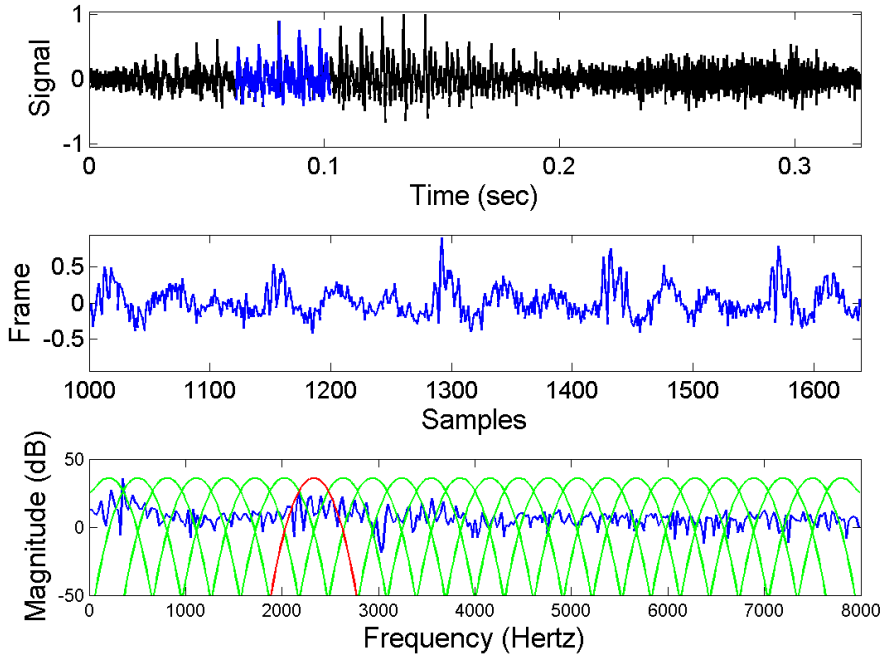
Attention Models: Bad Example

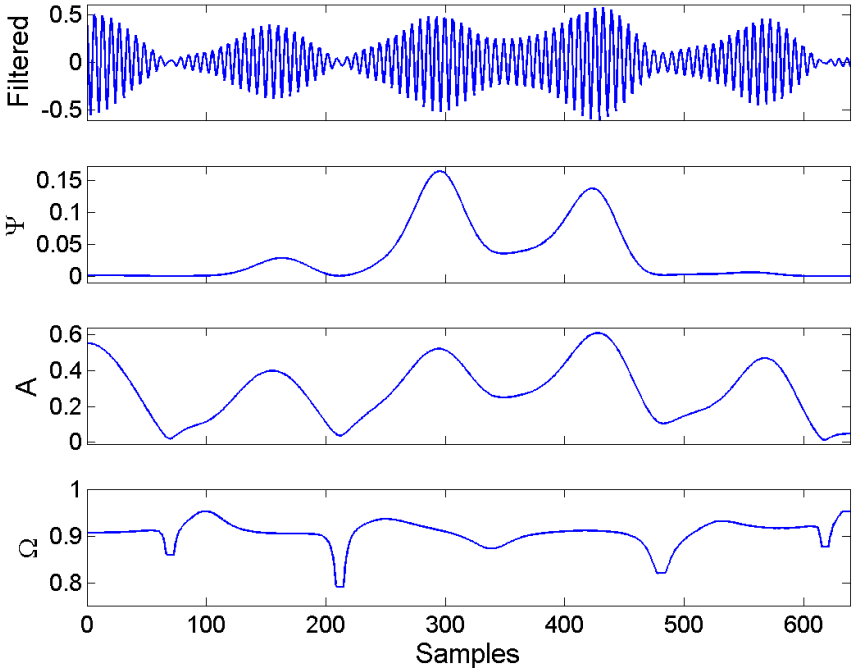


Attention Models and Saliency

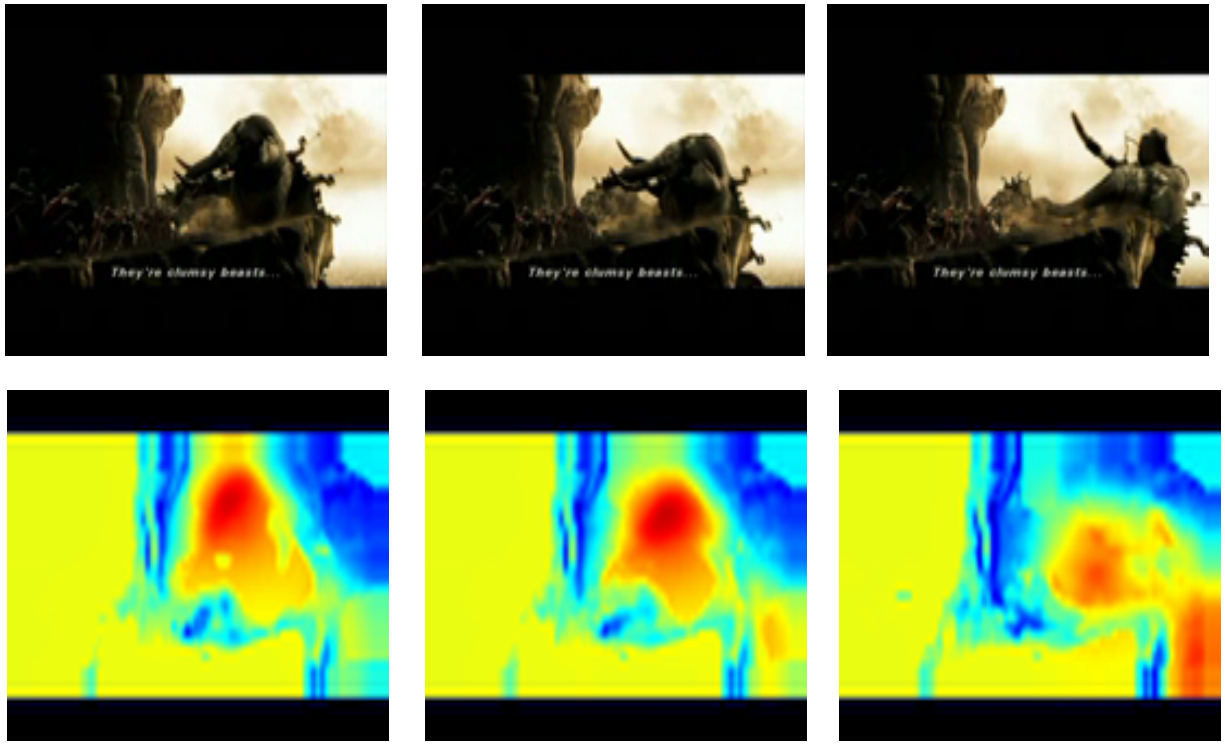
* Attention model of video streams
* Saliency measures:
  - Aural: energy of multi-frequency band features
  - Visual: multi-scale intensity, color and motion
  - Text: part of speech assignments
* Fusion on a single audio-visual-text saliency metric

Audio Saliency Features

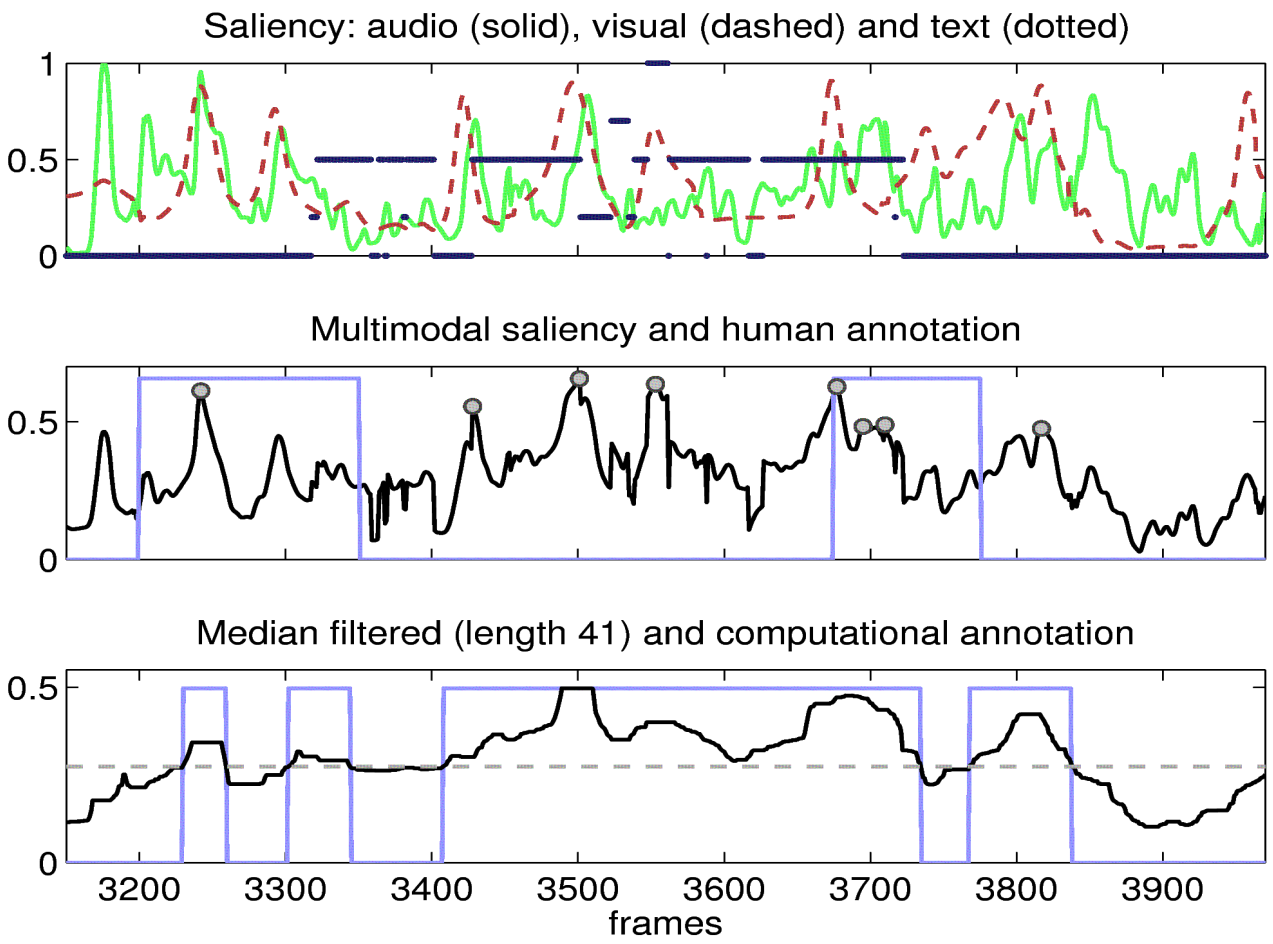




Visual Saliency



AVT Salience via Linear Fusion



Example: x2 compression

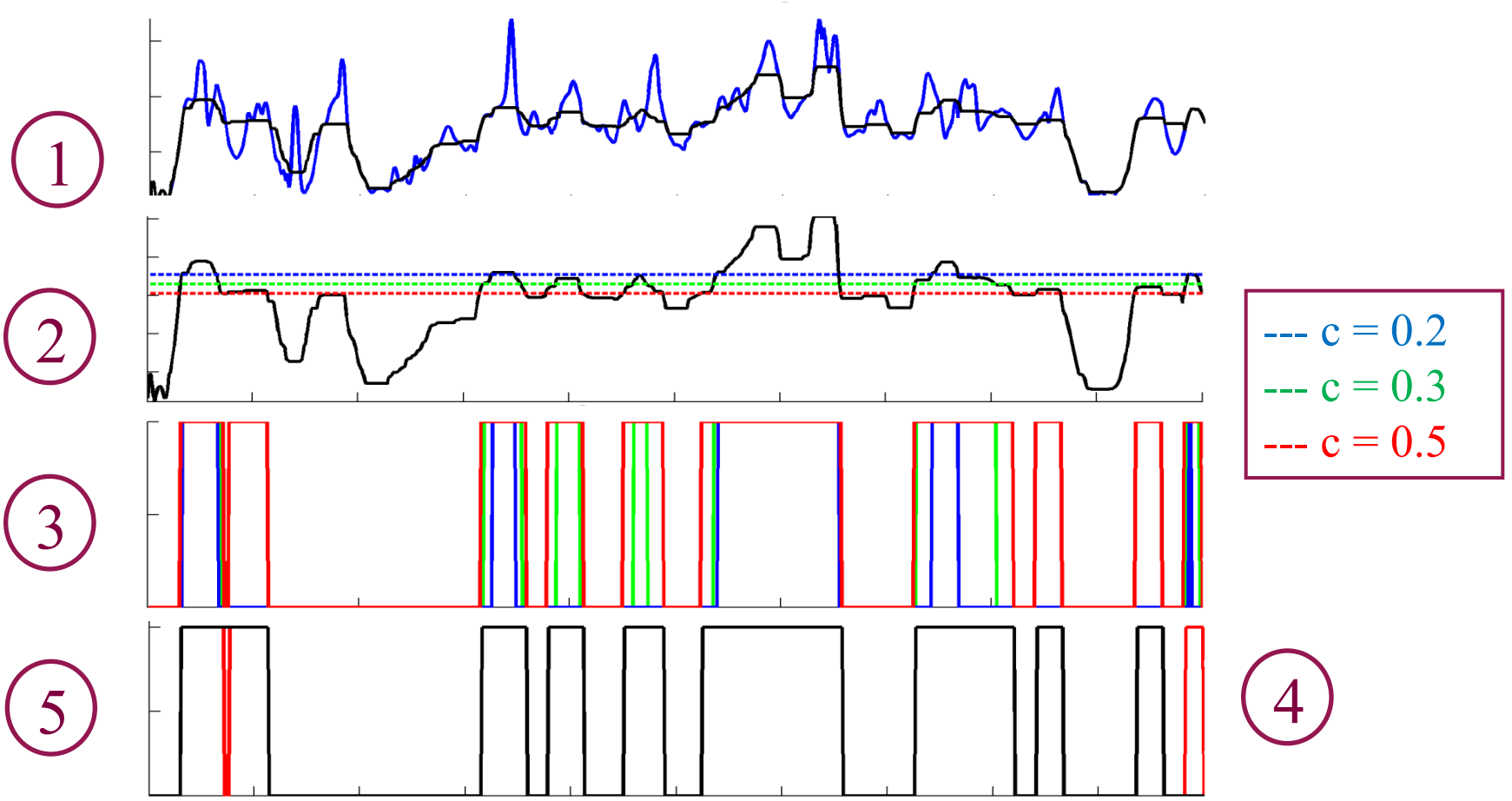
AV Key Frames: 300

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).

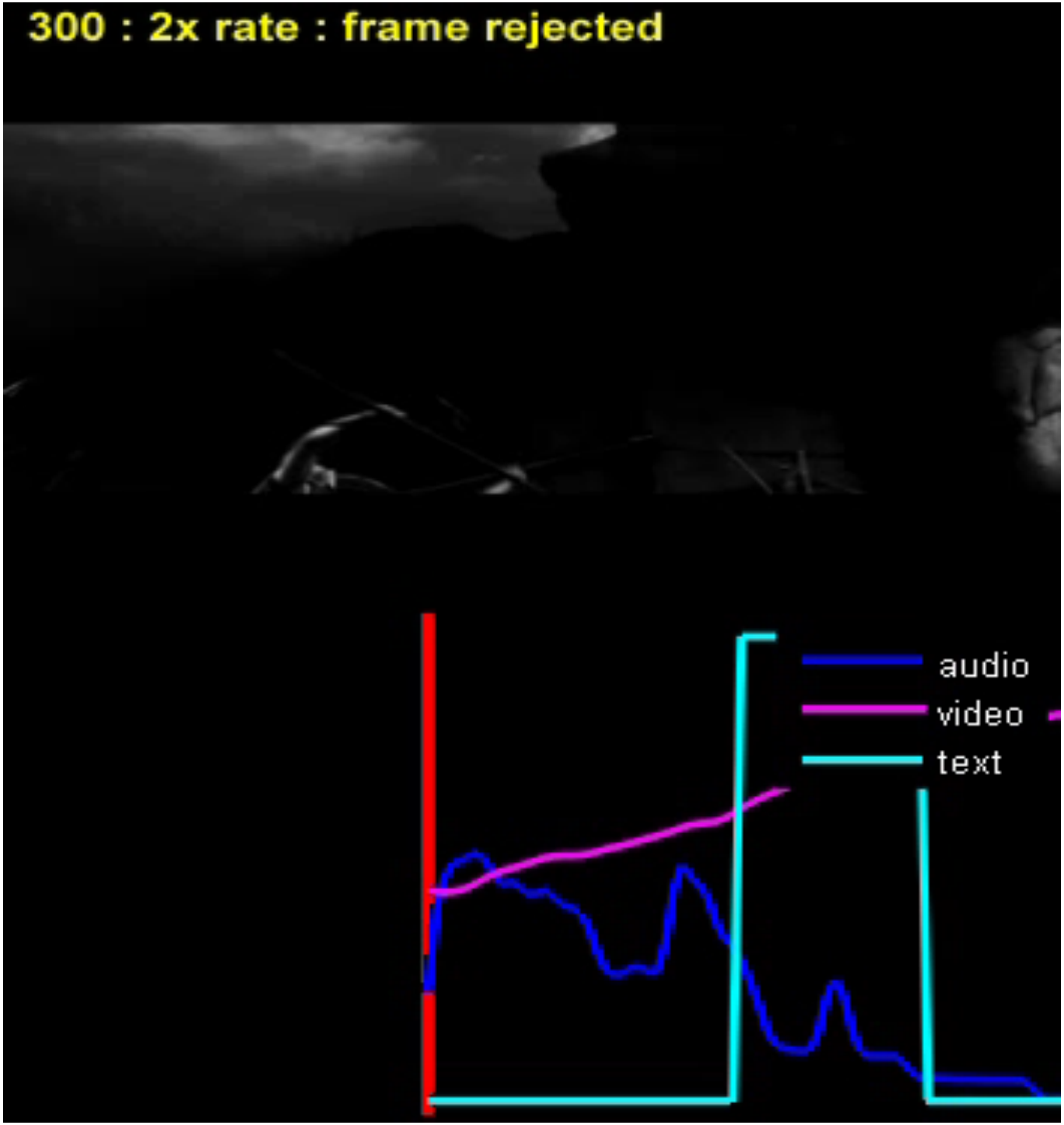
Movie Summarization Algorithm (2)



Summary annotated with AVT Saliency

Grey – Rejected

Color- Accepted in summary



Discussion

- Low-level selectional attention can be modeled using
  - Low level feature detectors
  - Fusion of detectors across modalities
  - Can capture up to 95% of semantics
- Relevance for audio processing
  - Audio source separation
  - Event detection

Semantic Representations

List of Open Questions

1. How are concepts, features/properties, categories, actions represented?
2. How are concepts, properties, categories, actions combined (compositionally)?
3. How are judgements (classification/recognition decisions) achieved?
4. How is learning and inference (especially induction) achieved?

Answers should fit evidence by psychology and neurocognition!

Three Solutions

- Symbolic
  - cognition is a Turing machine
  - computation is symbol manipulation
  - rule-based, deterministic (typically)
- Associationism, especially, connectionism (ANNs)
  - brain is a neural network
  - computation is activation/weight propagation
  - example-based, statistical, unstructured (typically)
- Conceptual
  - intermediate between symbolic and connectionist
  - concepts are represented as well-behaved (sub-)spaces
  - computation tools: similarity, operators, transformations
  - hierarchical, semi-structured

Properties of the Three Approaches

Property	Symbolic	Conceptual	Connectionist
cognitive speed	very slow	slow	fast
machine speed	very fast	pretty fast	fast
cognitive accuracy	good	good	decent
machine accuracy	decent	good	good
dimensionality	high	low	high
representation	flat	hierarchical	distributed
interpretability	excellent	good	low
determinism	high	medium	low
reasoning (all data)	good	good	decent
compositionality	good	good	decent
induction/learning	poor	excellent	average

Properties of the Three Approaches

- Symbolic
  - Good for high-level cognitive computations (math)
  - Poor generalization power
  - Too expensive and slow for most cognitive purposes
- Conceptual
  - Excellent generalization power (intuition, physics)
  - Good for induction and learning; geometric properties (hierarchy, low dim., convex) guarantee quick convergence
  - Properties and actions defined as operators/translations
  - Still too slow for some survival-dependent decisions
- Connectionist (machine learning)
  - General-purpose, extremely fast and decently accurate
  - Computational sort-cuts create cognitive biases
  - Poor generalizability power due to high dimensionality and lack of crisp semantic representation

Main approaches of lexical semantics

- Word are associated with feature vectors
  - crisp, parsimonious representation of semantics
- Distributional semantic models (DSMs)
  - Semantic information extracted from word frequencies
  - Estimate co-occurence counts of word pairs or triplets
  - Estimate statistics of word context vectors
- Semantic networks
  - discovery of new relations via systematic co-variation
  - robust estimates – smoothing corpus statistics over network
  - rapid language acquisition

Representation Learning

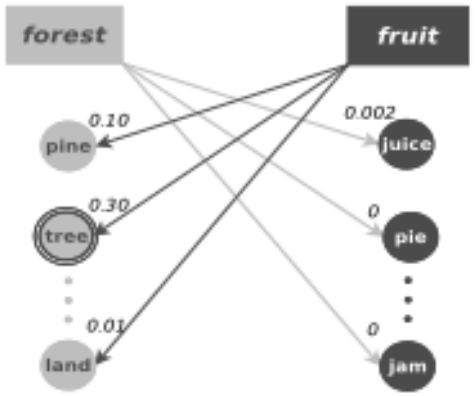
- Properties of a classifier with good generalization properties [Bengio et al 2013]:
  - Low-dimensionality/Sparseness
  - Distributed representations/hierarchy
    - Depth and abstraction
  - Shared factors across tasks
- Examples: auto-encoders, manifolds, deep neural nets …
- How to induce these properties in your classifiers:
  - Include as regularization term in training classifier criterion
  - Include properties directly in classifier design
  - Go deep and pray (dirty neural net tricks)

Proposed semantic similarity two-tier system

- Unifies the three approaches
- Fuzzy vs explicit semantic relations
- Word senses vs words vs concepts
- A two tier system
  - An associative network backbone
  - Semantic relations defined as operations on network neighborhoods (cliques)
- Consistent with system 1 vs system 2 view
- Furthermore we believe that the
  - underlying network consists of word senses, and
  - is a low dimensional semi-metric space

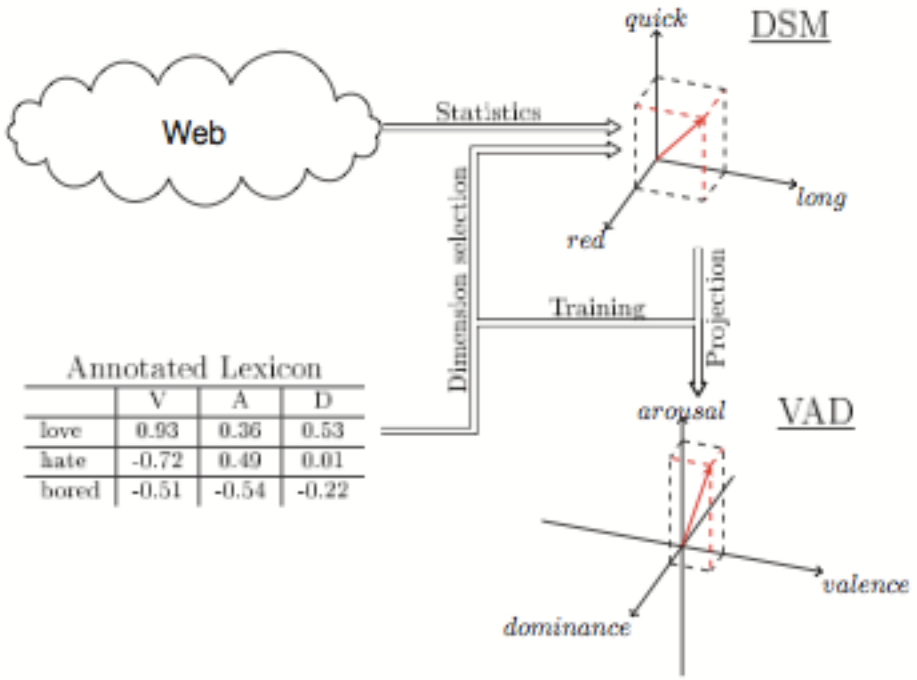
Neighborhood-based Similarity Metrics: M_n

M_n metric: maximum similarity of neighborhoods



- Motivated by maximum sense similarity assumption
  - Neighbors are semantic features denoting senses
  - Similarity of two closest senses
- Select max. similarity: $M_n(\text{“forest”}, \text{“fruit”}) = 0.30$

Computations are mappings between layers



Annotated Lexicon

	V	A	D
love	0.93	0.36	0.53
hate	-0.72	0.49	0.01
bored	-0.51	-0.54	-0.22

Our lexicon expansion method

Expansion of [Turney and Littman, '02].
Assumption: the valence of a word can be expressed as a linear combination of its semantic similarities to a set of seed words and their valence ratings:

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i \, v(w_i) \, d(w_i, w_j), \qquad (1)$$

- w_j : the wanted word
- $w_1 ... w_N$: seed words
- $v(w_i)$: valence rating of word w_i
- a_i : weight assigned to seed w_i
- $d(w_i, w_j)$: measure of semantic similarity between words w_i and w_j

Grand Challenges

Detection-based Audio Processing

- …

Saliency-driven Multimedia Processing

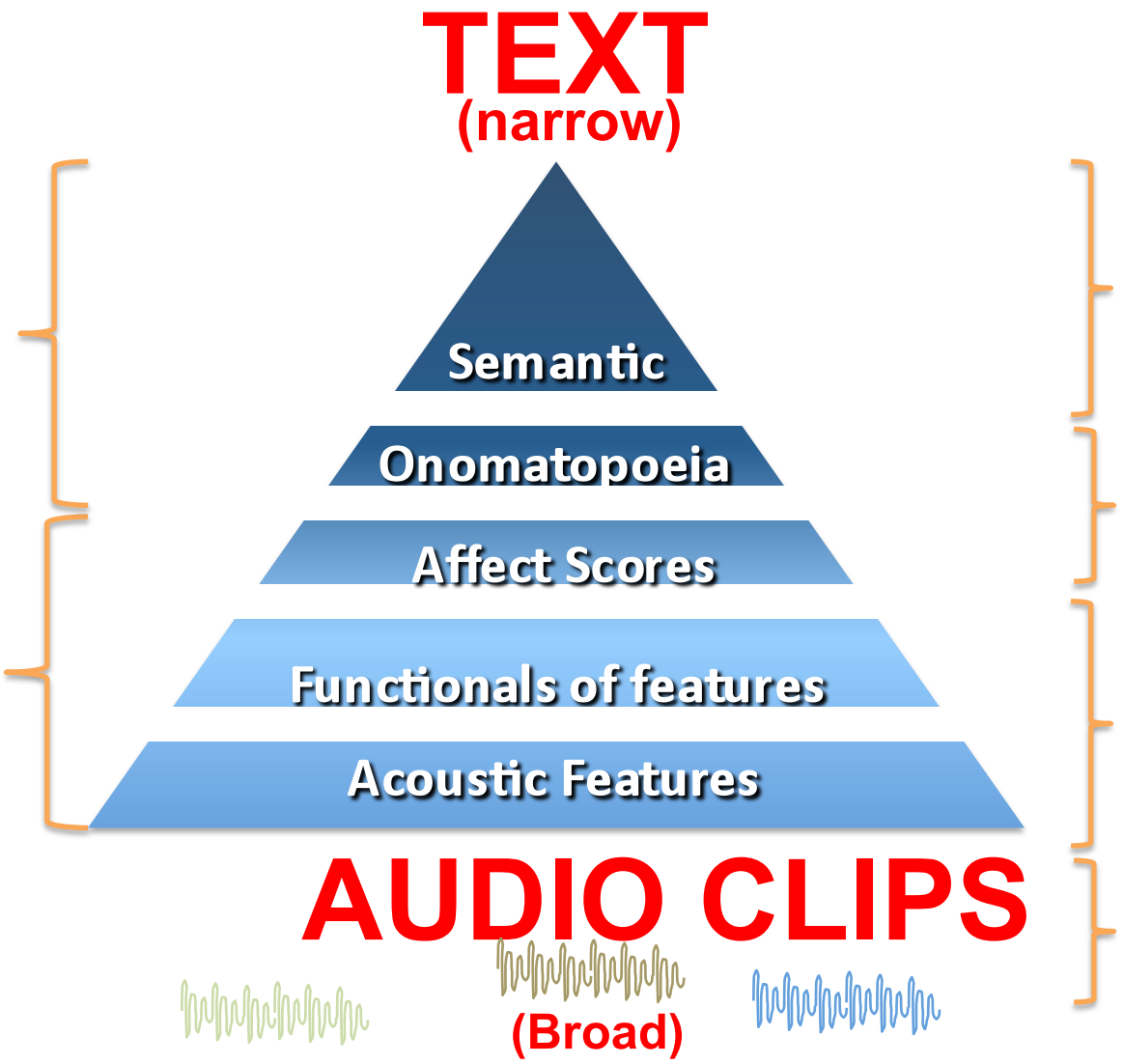
- …

Representation Models for Multimedia

- Similarity is the main building block
  - 3 types: similarity w. internal semantic representation, self-similarity over time, similarity in context (biases by world/internal view)
  - Associative network is layer 1 – all computations use this basic representation
- Detectors live in low-dimensional spaces with good geometric properties ("metric")
- Features are labels, labels are features
- Features/labels are organized hierarchically (multiple layers from specific to general, i.e., abstraction)

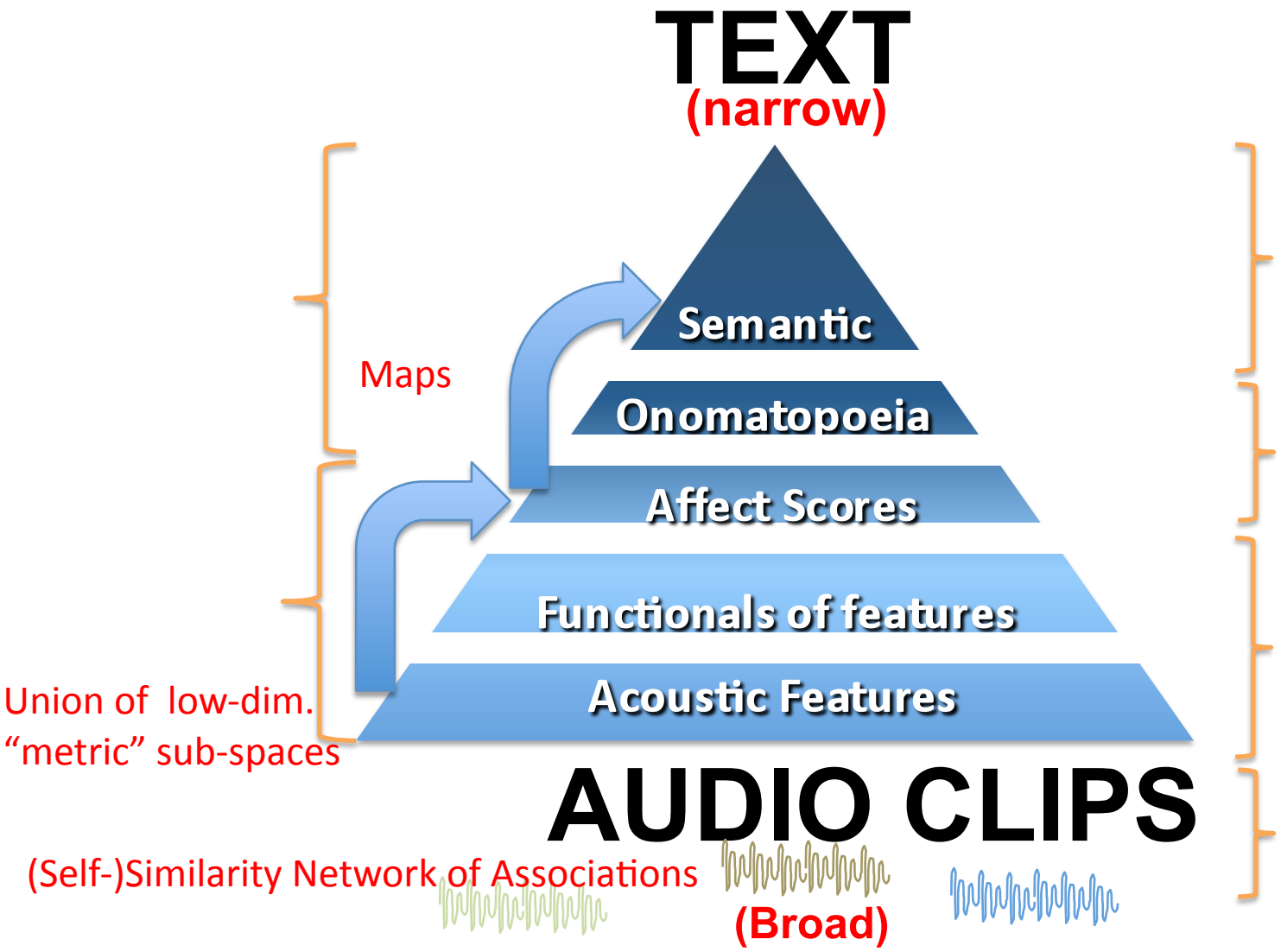
Descriptions of Sounds

[slide by Shiva Sundaram]



Descriptions of Sounds

[original slide by Shiva Sundaram]



Our Timeline

- Unexpectedly good results on semantic similarity tasks using web data
- [E. Iosif, and A. Potamianos, "Unsupervised Semantic Similarity Computation Between Terms Using Web Documents,'' *IEEE Transactions on Knowledge and Data Engineering*, Nov. 2010]
  - Lucky enough to: 1) work on a semantic similarity task,
    2) directly modeling human cognition
- Goal: reduce web query complexity from quadratic to linear
  [E. Iosif, and A. Potamianos, "Similarity Computation Using Semantic Networks Created From Web-Harvested Data", *Natural Language Engineering*, 2013]
  - Lucky enough not to stop at good initial performance
- Realization:
  - generalization power is in the semantic representation/network
  - multi-tier models: associative network is the 1st tier
- Cognitive science literature [P. Gardenfors, *Conceptual Spaces,* 2000]
  - Low-dimensional "metric" sub-spaces (good geometric properties)
  - Maps and operators defined in this space
- Combine experience from machine learning to come up with a general model