Multimodal Human State &Trait Recognition: Quo plures, eo feliciores?

Björn Schuller @ TGMIS 2014

Imperial College London

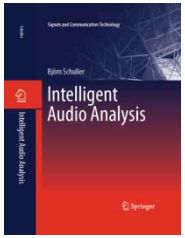
audEERING intelligent Audio Engineering

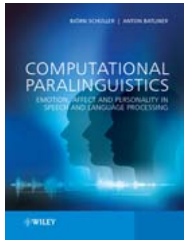
UNIVERSITÄT PASSAU

Introduction

Presenter.

> 430 publications.
> 7100 citations.
h-index = 41.
President AAAC.
Editor in Chief IEEE T-AC.



Imperial College London

Diploma	PhD	Habilitation	
1999	*2006*	*2012*	

Research Assistant	Lecturer	Visitor	Professor
2000	*2006*	*2009*	*2012*

Team & Projects.

>5 Mio EUR funding.
ERC Starting Grant.
ERC Advanced Grant Participant.

DAIMLER SEMAINE ASC-Inclus

SIEMENS VDO Automotive

SEVENTH FRAMEWORK PROGRAMME

Interactive Emotion G

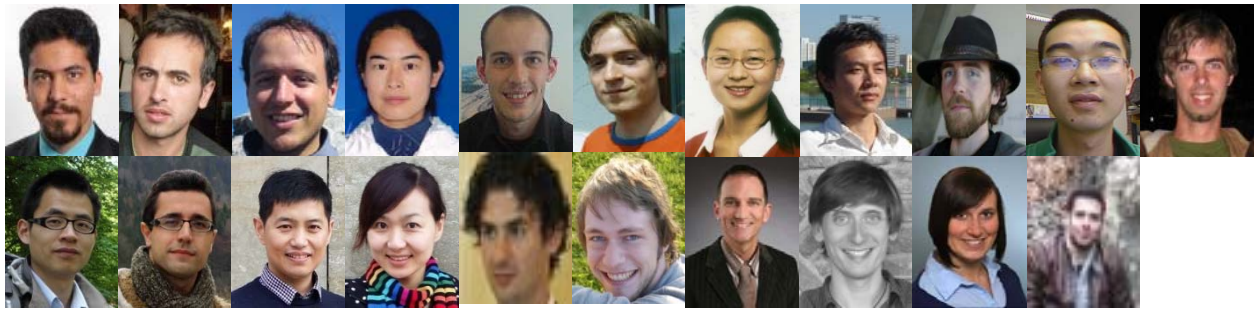
W3C WORLD WIDE WEB consortium

Continenti

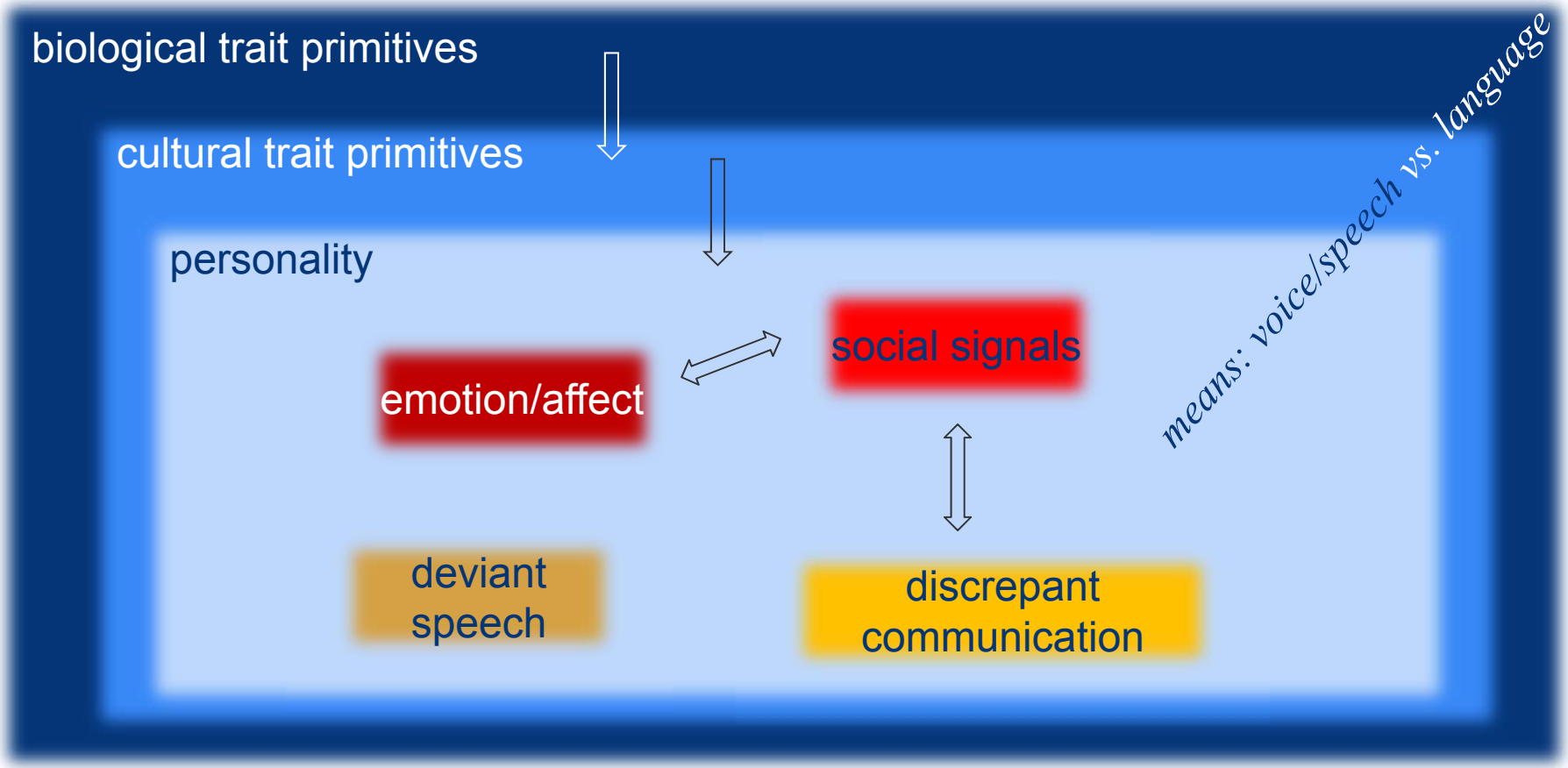
Deutsche Telekom Laboratories

TOYOTA

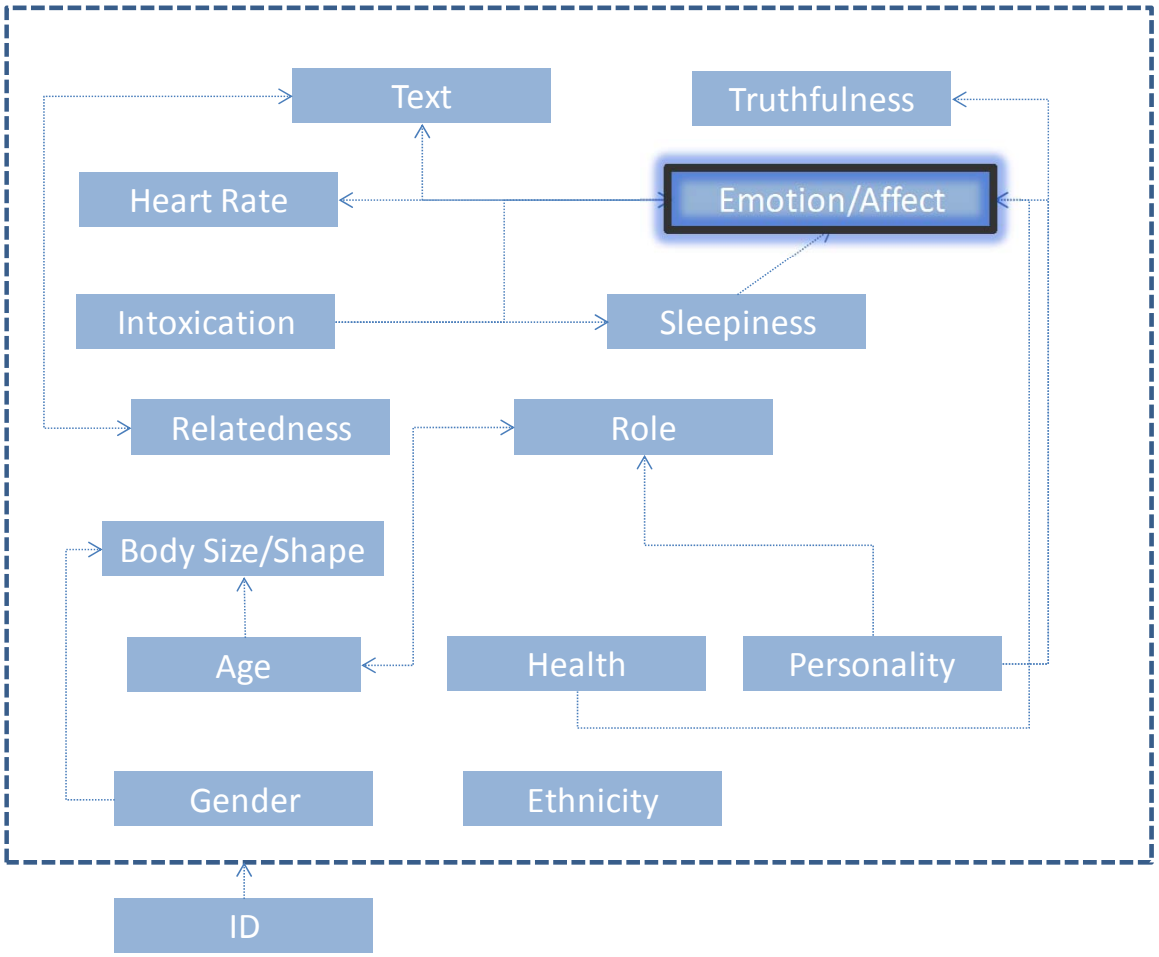
DF



States & Traits.



Holism.

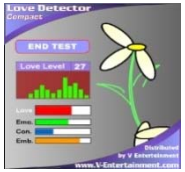

Text
Truthfulness
Heart Rate
Emotion/Affect
Intoxication
Sleepiness
Relatedness
Role
Body Size/Shape
Age
Health
Personality
Gender
Ethnicity
ID

Applications.

- HCI – Sensitive Artificial Listeners, attention analysis, call centres, car, …
- HRI – humanoid robotics
- Multimedia – video content representation, affective audio/video retrieval, coding
- Entertainment technology – gaming, arts
- Learning environments – episodic learning, coaching on paralinguistics
- Smart home – ambience control
- Monitoring – safety, assessment, customers
- Clinical & biomedical studies – speech disorders, Altzheimer, Parkinson, stress/pain monitoring, Autism-related assistive technology, Rett and Fragile X syndrome, sleep disorders

Products.

- **Multimodal Yet… ?**



Recent Cases

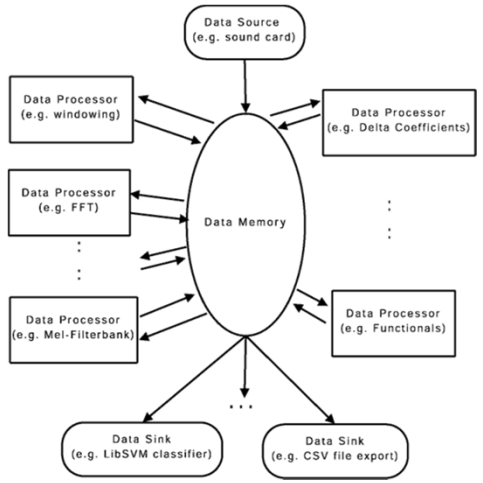
Feature Brute-Forcing.

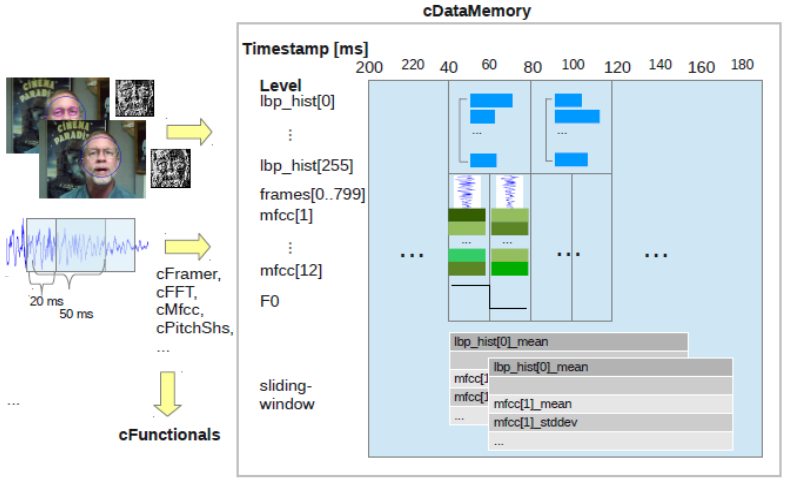
- **openSMILE 2.0: Audio/Visual + X**

Distributed Feature Extractor
(Android / C++)

Multithreading
Memory efficient

#features	RTF
10k	2%
500k	**3%**





"Recent Developments in openSMILE, the Open-Source Multimedia Feature Extractor", **ACM Multimedia**, 2013.
(2nd place ACM MM Open Source Software Competition in 2010 and 2013)

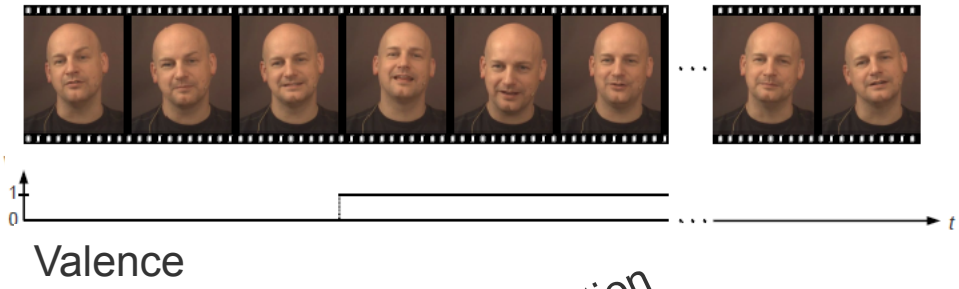
AVEC 2011.

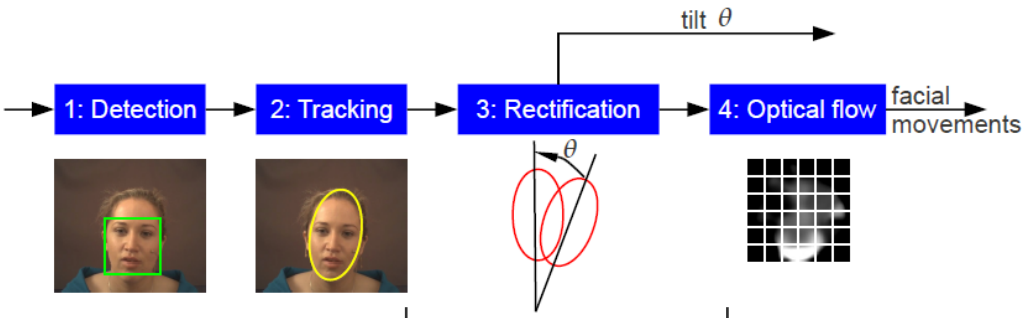
- **Audiovisual Emotion**

  AVEC 2011/12: 50,4k Turns

  L(G)BP, optical flow +

  acoustic + vocalisations + text

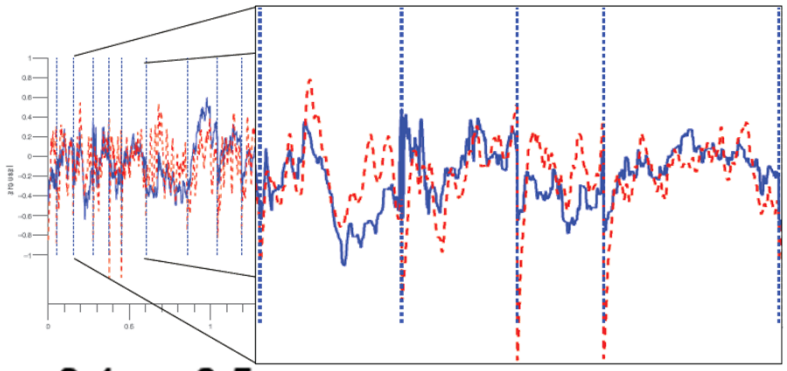




WA [%]	Activation	Expectation	Power	Valence
Audio	**71.2**	63.7	**62.2**	**70.2**
Video	48.6 (53.2)	**68.6**	57.9	69.6
All	70.3	64.1	57.5 (62.9)	69.6

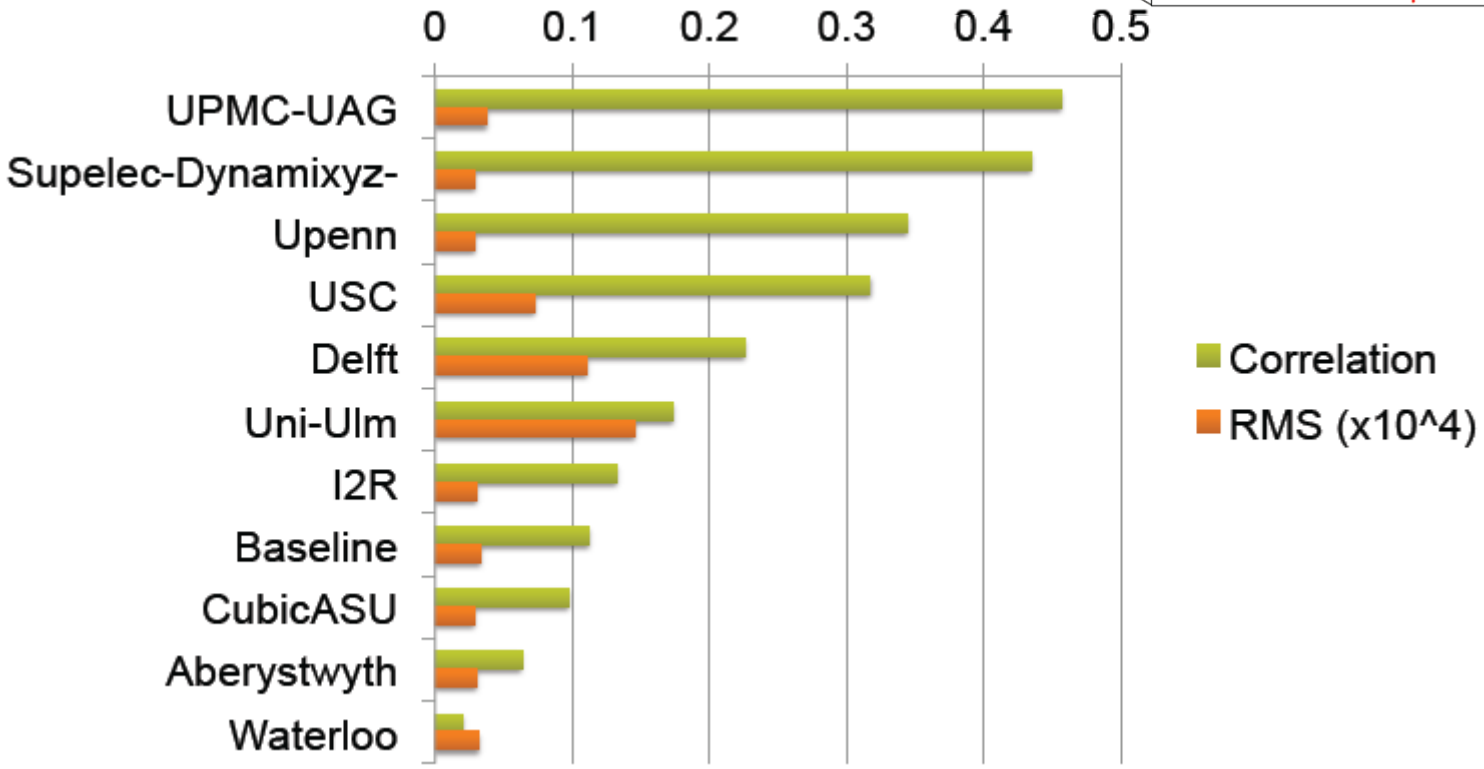
*"LSTM-Modeling of Continuous Emotions in an Audiovisual Affect Recognition Framework", **Image & Vision Computing Journal**, 31(2): 153-163, 2012.* ***(Best Result AVEC Challenge)***

AVEC 2012.

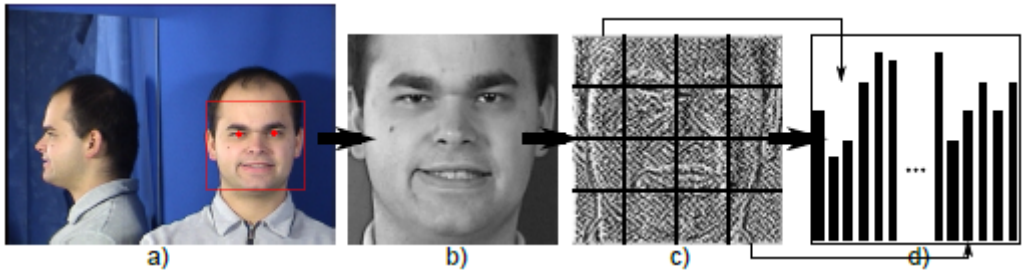


- **Fully Continuous Sub-Challenge**

  3 x 500,000 frame / label pairs



AVEC 13/14.



- **Data**

  Audio-visual depressive language corpus (AVDLC)
  340 video clips of subjects performing a HCI task,
  Total duration 240 hours, 292 subjects

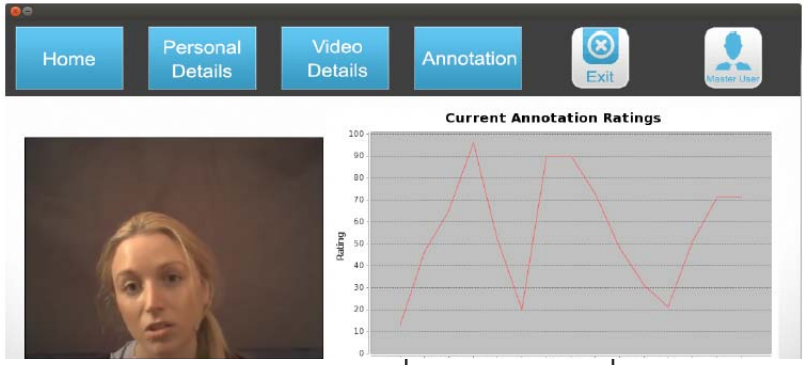
- **Depression / Affect Recognition**

  Beck Depression Inventory (BDI): 0 – 63

MAE	Depression
Winner	6.5

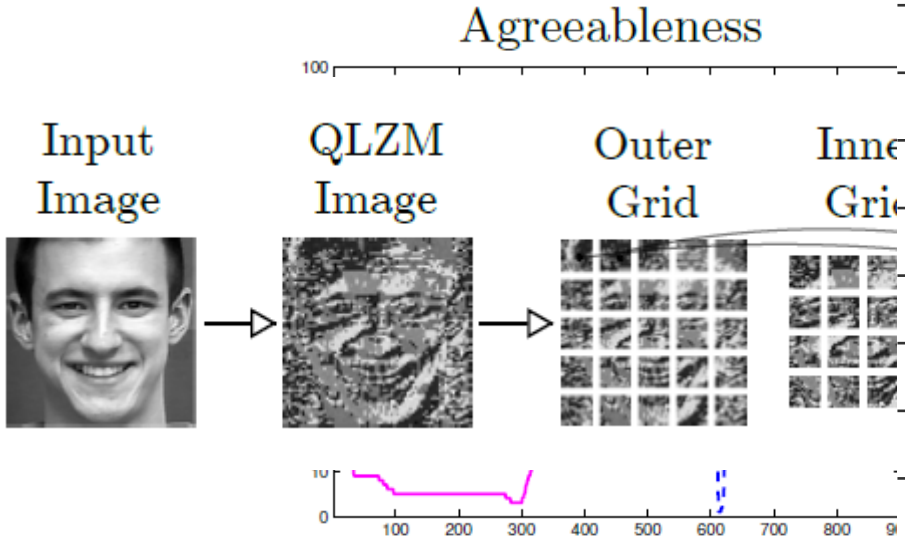
"AVEC 2013 - The Continuous Audio/Visual Emotion and Depression Recognition Challenge", ***ACM Multimedia****, 2013.*

MAPTRAITS.



- **Data**

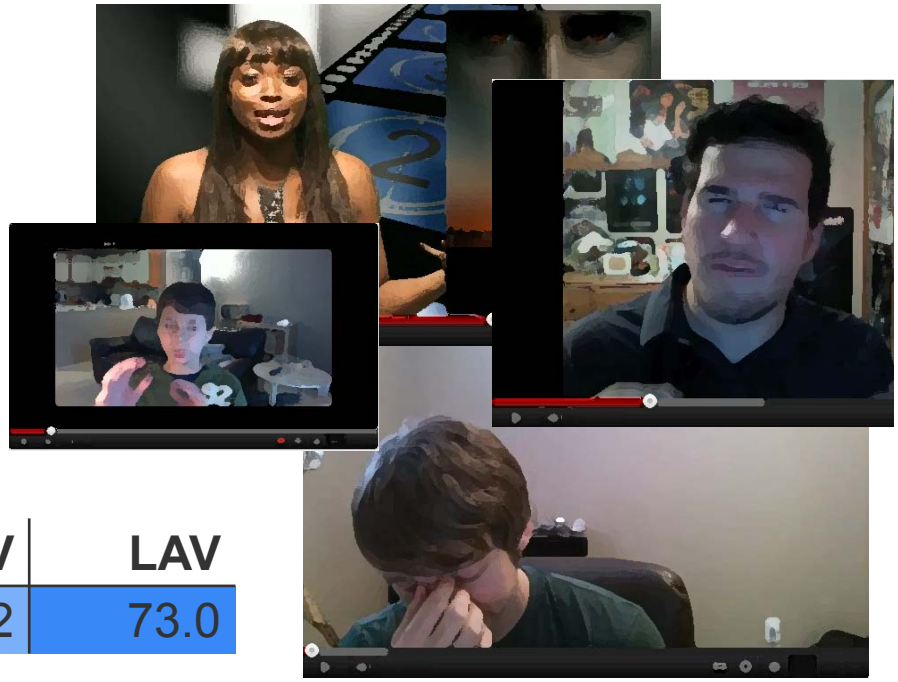
  SEMAINE:
  44 clips, 11 subjects, 6 raters
  Late fusion: CC+ likability, agreeablen.



MSE	Audio	Video
Openness	.53	.39
Conscientious.	.51	.28
Extroversion	.66	.38
Agreeableness	.50	.41
Neuroticism	.77	.35
Engagement	.65	.45
Likeability	.56	.37
Facial Attractiv.	-	.34
Vocal Attractiv.	.41	-

*"MAPTRAITS 2014: The First Audio/Visual Mapping Personality Traits Challenge", **ACM ICMI**, 2014.*

Sentiment.



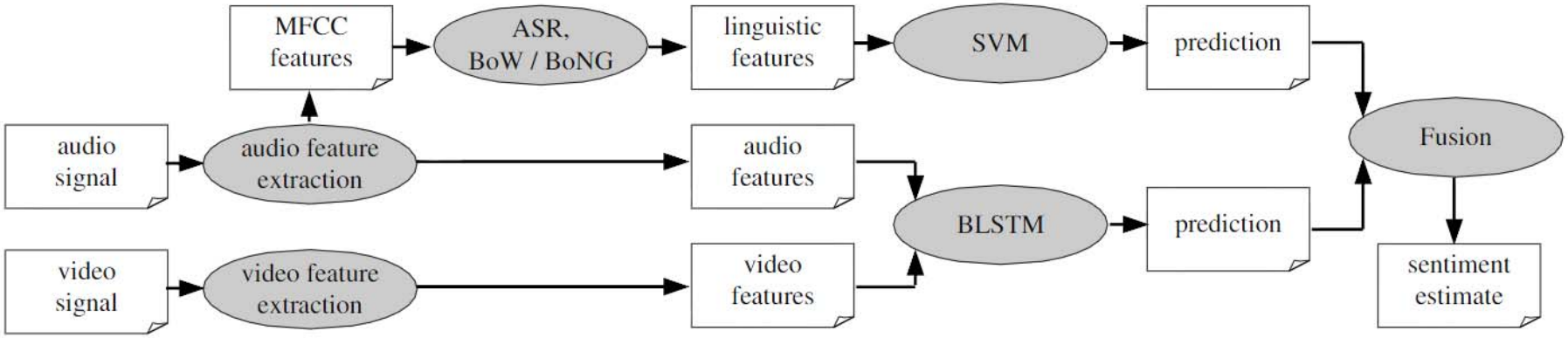
- **Audiovisual Sentiment Polarity**

  Multi-Modal Movie Opinion Database

  370 videos: YouTube & ExpoTV

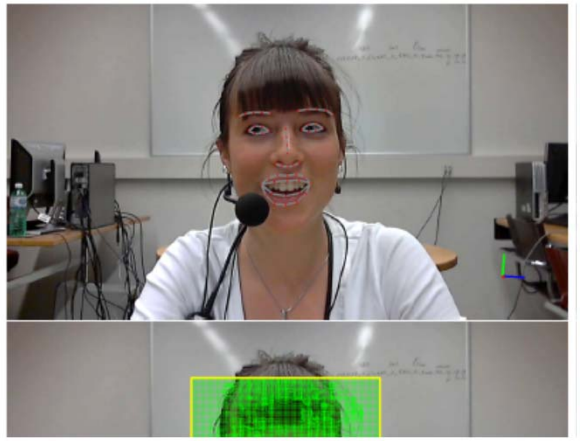
  Amateur Movie Reviews

UA [%]	Audio	Video	AV	LAV
Polarity	64.4	61.2	66.2	73.0



"YouTube Movie Reviews: In, Cross, and Open-domain Sentiment Analysis in an Audiovisual Context"
***IEEE Intelligent Systems Magazine**, 2013.*

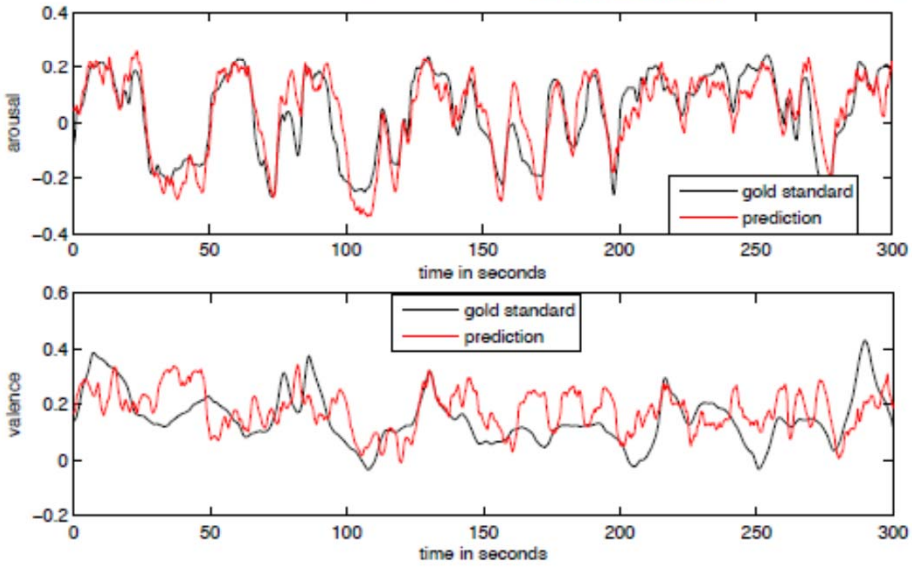
Emotion.



- **Audiovisual + Physiology (ECG/EDA)**
  - RECOLA Database
  - 27 subjects, 5 min each, 6 raters
  - Collaborative Interactions
  - ComParE, 15 AUs, 3x Headpose, 28xphysio

CC [%]	Arous	Valen
Audio	78.8	34.3
Video	42.7	43.1
All	**80.4**	**52.8**

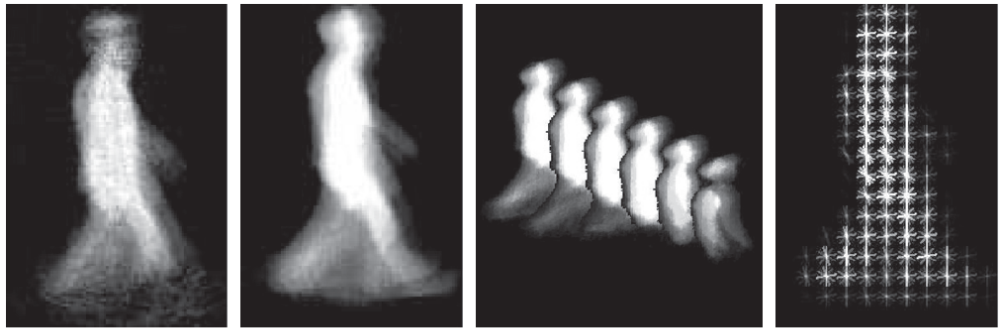
  - Fusion better late
  - Valence longer window sizes



"Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data"
***Pattern Recognition Letters**, 2014.*

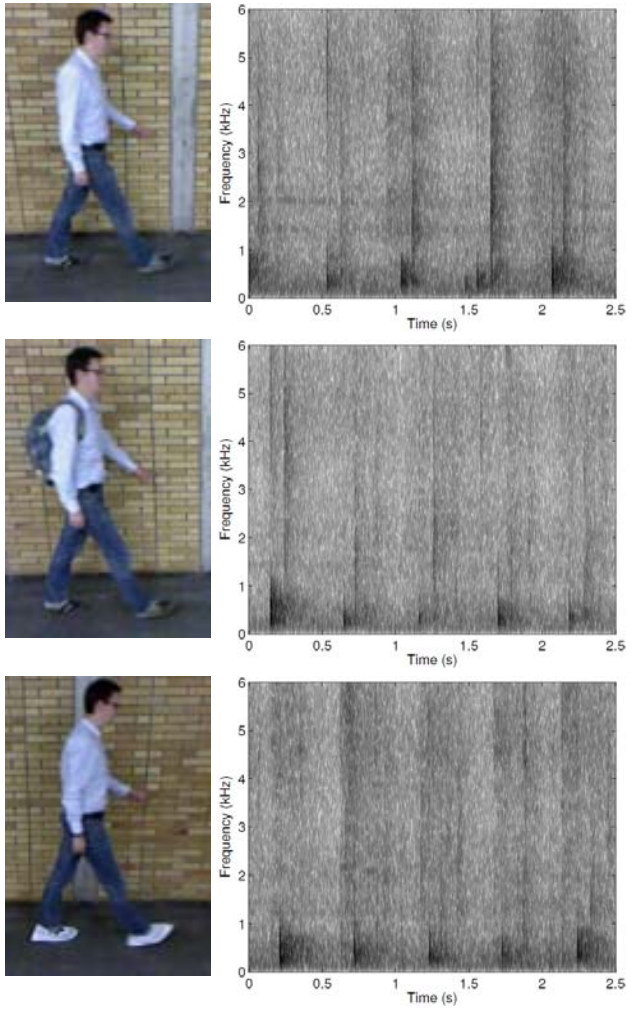
Biometrics.

- **Biometrics from Walking Patterns**
  GAID database (Kinect): 305 subjects
  Words: LIWC, Video: LBP



(a) GEI (b) depth-GEI (c) GEV (d) DGHEI

UA [%]	Words	Audio	Video	WAV
Age	64.6	62.7	71.0	72.9
Sex	75.4	93.2	81.4	-
Race	67.1	52.3	70.3	73.4



"The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits", ***Journal of Visual Communication and Image Representation****, 2013.*
"Speaker Trait Characterization in Web Videos: Uniting Speech, Language, and Facial Features" ***IEEE ICASSP****, 2013.*

Attention.

- **Attention Recognition**
  - Audi A6, real street
  - Audi Multimedia System
  - On-board computing
  - Off-board analysis
  - 30 drivers (23-59 years)
  - 8 typical interaction tasks
  - CAN-bus + inner camera

"On-line Driver Distraction Detection using Long Short-Term Memory", ***IEEE Transactions on Intelligent Transportation Systems,*** *12(2): 574-582, 2011.*

- **Testing**
  - Driver-independent

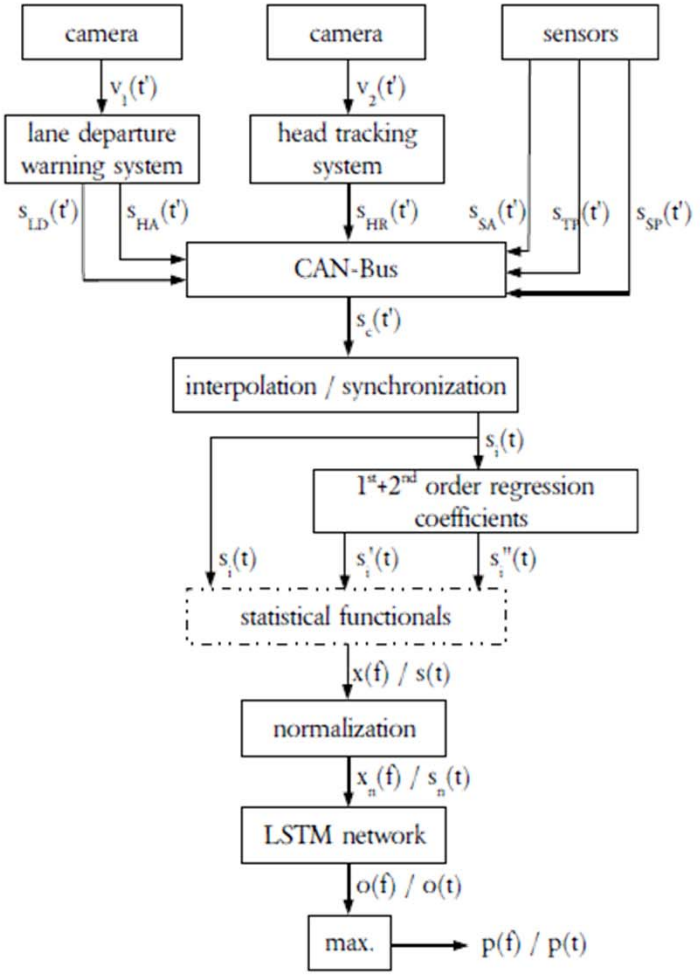
	UA [%]
hi/low	**95.0**
hi/med/low	**70.2**

Attention.

- **Features (CAN-bus + Sensors)**

Steeringwheel angle	(25.0%)
Pedal position	(3.1%)
Speed	(5.2%)
Driving angle	(7.9%)
Lateral deviation	(5.7%)
Head rotation	(53.1%)

LLD x 3 and 55 Functionals
Extremes (7), Regression (9), Means (7)
Percentiles (6), Peaks (4), other (22)



Discussion

Discussion

- **More = more?**
  - Often not…
  - Often not very significant
  - By performance on datasets
  - Lack of usability studies in the real world
- **Weak/Strong Modalities?**
  - Often in the literature “favourite modality” + “other one(s)”
- **Same Data**
  - Usually trained across modalities for same data
  - What if not?

Discussion

- **Complementarity**
  Video/linguistics for valence, audio/physio for arousal, etc.
  But hardly approaches that *explicitly* use this fact

- **Fail-Safer?**
  Occlusions, noise, non-presence such as silence, etc,.
  But few results/tests with actual drop-outs
  What about culture / languages?

- **Ground Truth**
  Made for which modality/ies?

- **Mixed Fusion Approaches**
  Database dependent (RECOLA vs AVIC, etc.)?

Discussion

- **It's about Timing**
  - Per modality
  - Per state/trait
- **What about other factors?**
  - Speed, memory, privacy, confidence measures, distribution, …
- **Transfer Learning**
  - Possible across modalities?
- **Even More?**
  - Modalities (e.g., touch, smartphone sensors, etc.)
  - States & traits (many not approached multimodal, yet)

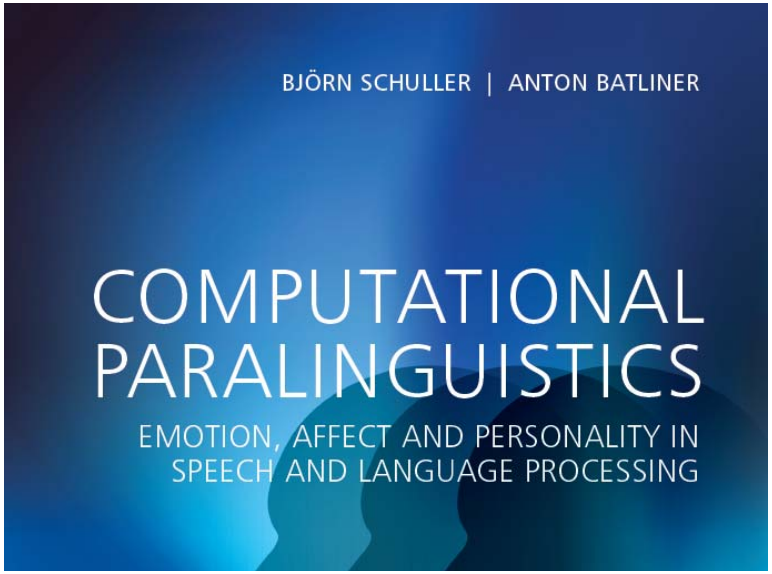
Speech.

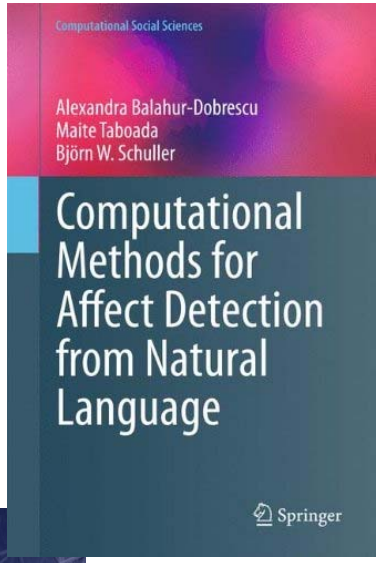
*"The Computational Paralinguistics Challenge", **IEEE Signal Processing Magazine**, 29(4): 2-6, 2012.*

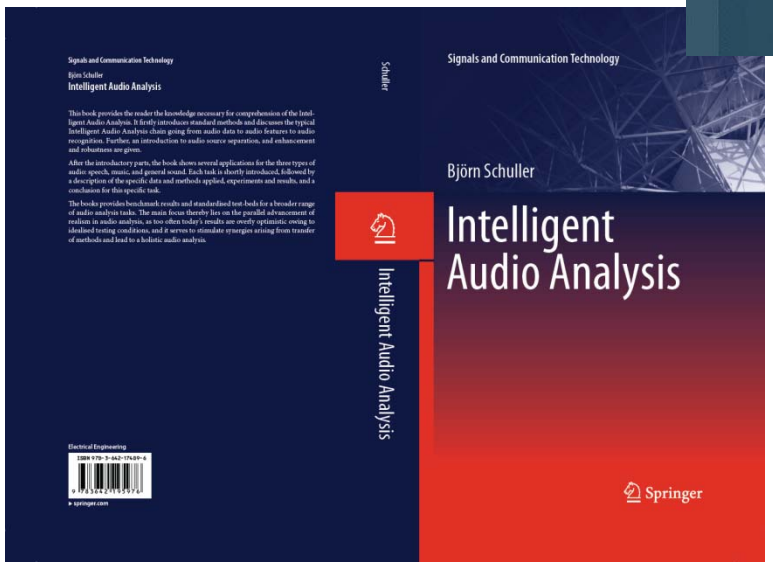
		# Classes	UAR/*UAAUC/+CC [%]
2014	Cognitive Load	3	68.9
	Physical Load	2	77.5
2013	Social Signals	2x2	92.7*
	Conflict	2	85.9
	Emotion	12	46.1
	Autism	4	69.4
2012	Personality	5x2	70.4
	Likability	2	68.7
	Intelligibility	2	76.8
2011	Intoxication	2	72.2
	Sleepiness	2	72.5
2010	Age	4	53.6
	Gender	3	85.7
	Interest	[-1,1]	42.8+
2009	Emotion	5	44.0
	Negativity	2	71.2

Activities

Coaching & Conversion







teşekkür ederim!
vielen Dank!

www.openaudio.eu

Abstract

Human state and trait recognition plays an ever increasing role in today's intelligent user interfaces lending them social competence for improved naturalness of the interaction. Obviously, such automatic assessment of user characteristics including emotion, personality or cognitive and physical load to name just a few is challenging. To ease this fact, it is broadly believed that a multimodal approach to the goal is beneficial. Here, we touch upon the question often arising when it comes to consideration of multiple modalities in computer-based human behavior analysis: the more the merrier? The modalities considered comprise the "usual suspects", namely speech, facial expression, and physiology alongside less typical candidates. Synergies are highlighted such as complementarity in view of emotion or personality primitives alongside arising problems of multimodal fusion. Examples include such from a series of recent public research competitions co-organized by the presenter.