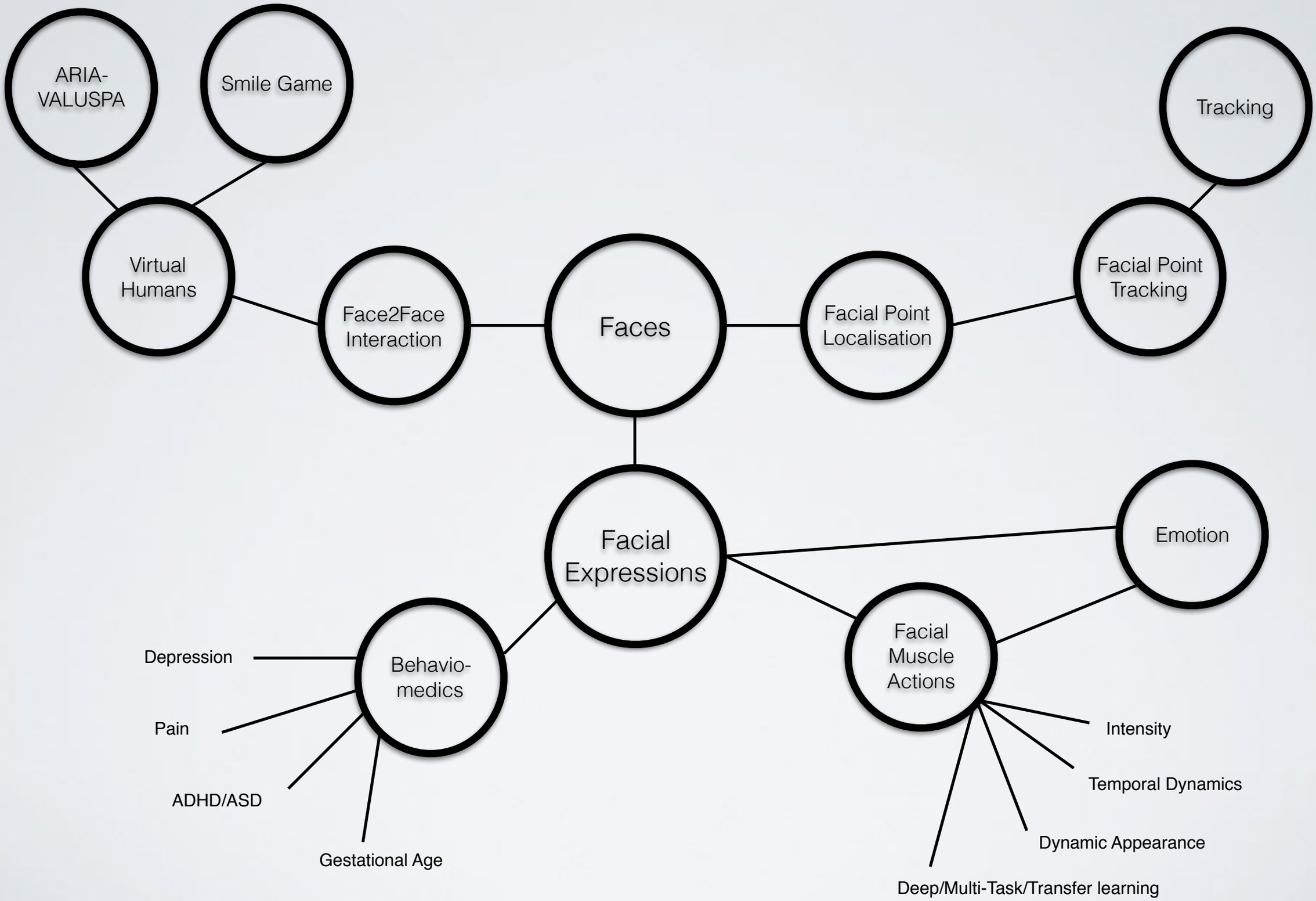
The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

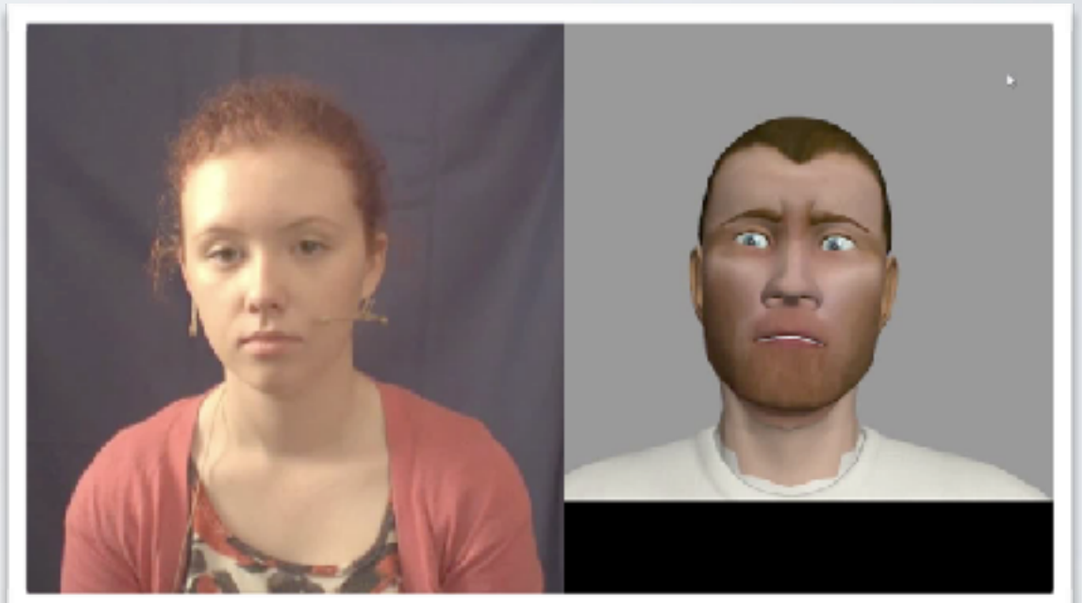
Automatic Human Behaviour Understanding

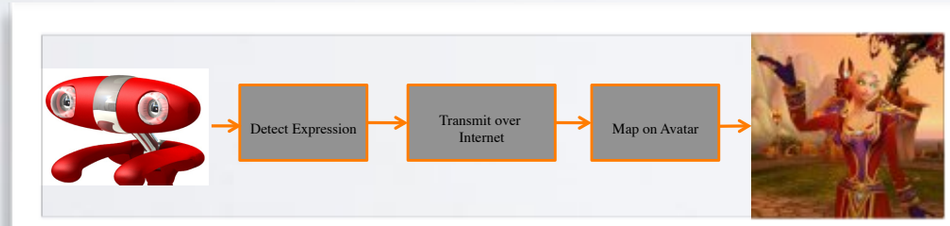
Facial Expression Recognition and its Application to Monitoring Medical Conditions

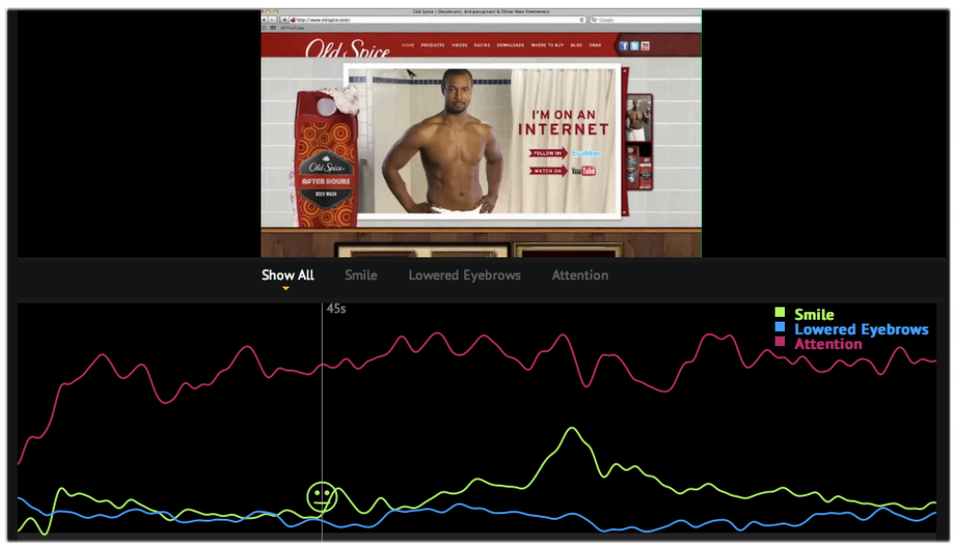
Michel Valstar
University of Nottingham
School of Computer Science
Mixed Reality/Computer Vision Lab

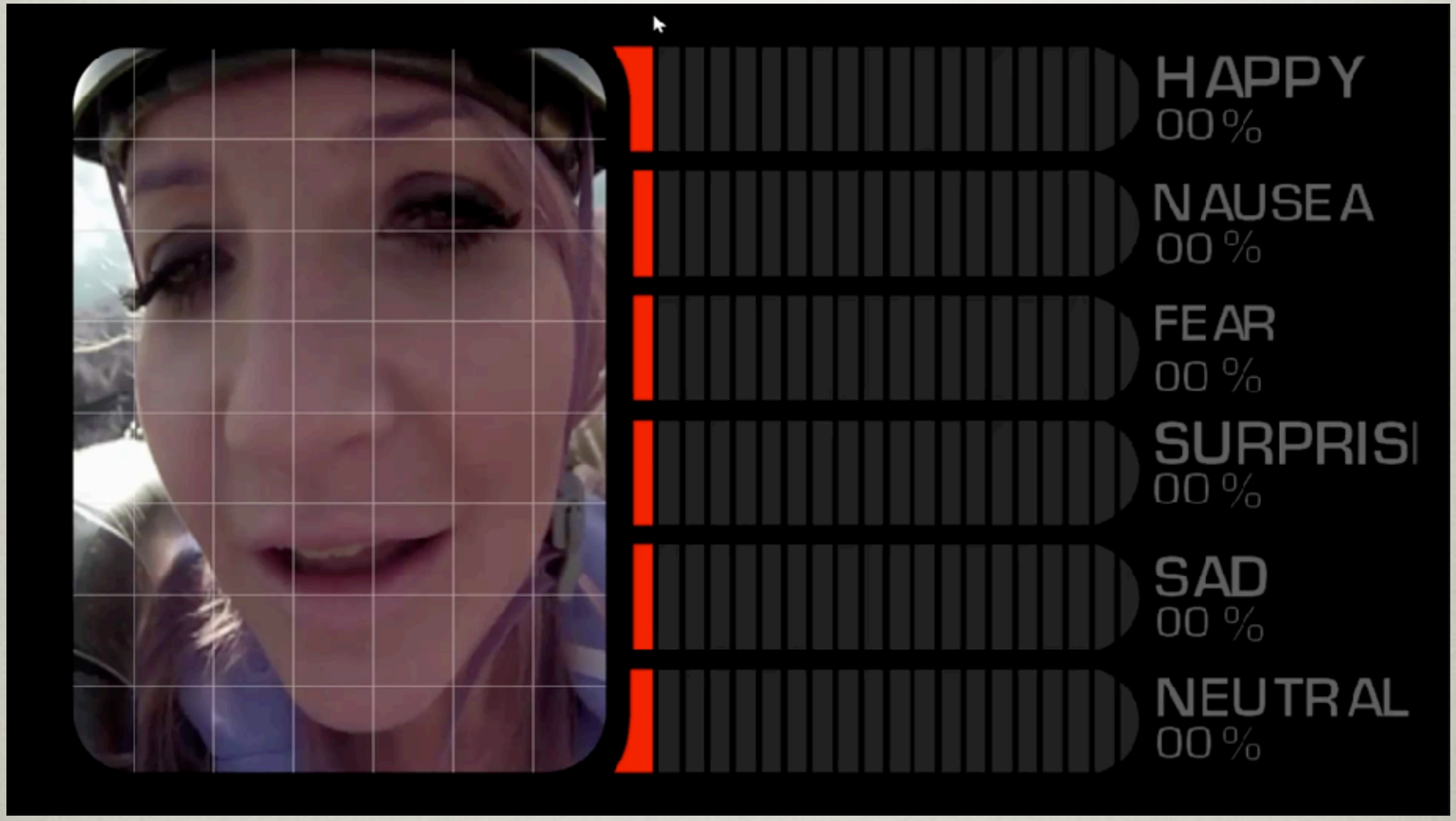

ARIA-VALUSPA
Smile Game
Virtual Humans
Face2Face Interaction
Faces
Facial Point Localisation
Facial Point Tracking
Tracking
Facial Expressions
Emotion
Facial Muscle Actions
Behavio-medics
Depression
Pain
ADHD/ASD
Gestational Age
Intensity
Temporal Dynamics
Dynamic Appearance
Deep/Multi-Task/Transfer learning

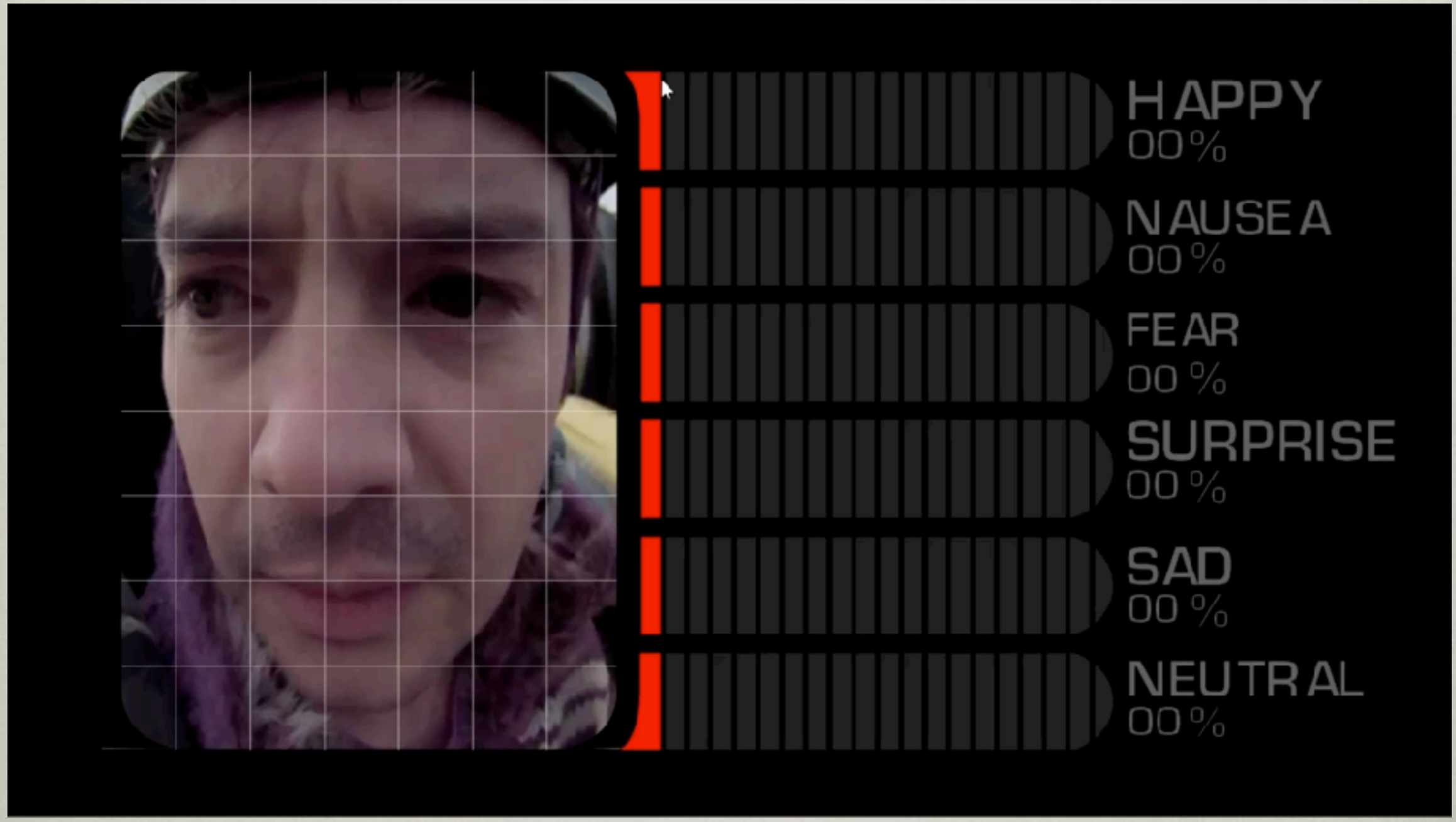
OPPORTUNITIES





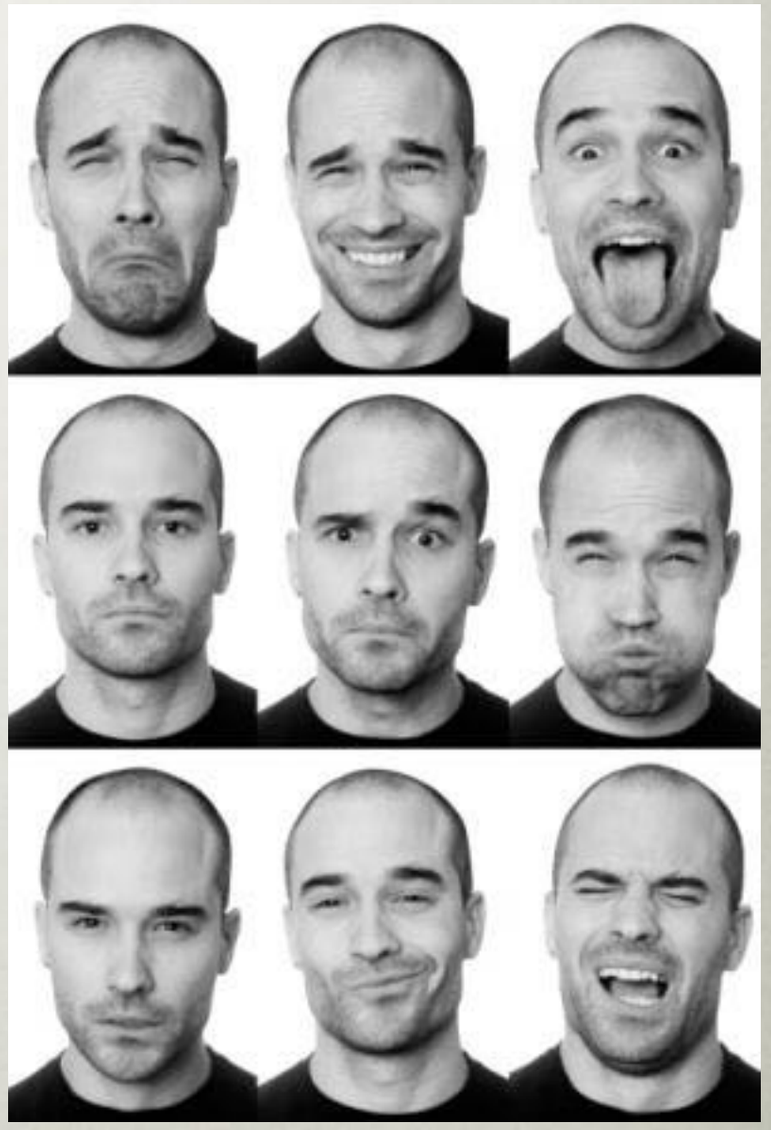



HAPPY
00%
NAUSEA
00%
FEAR
00%
SURPRISE
00%
SAD
00%
NEUTRAL
00%


HAPPY
00%
NAUSEA
00%
FEAR
00%
SURPRISE
00%
SAD
00%
NEUTRAL
00%

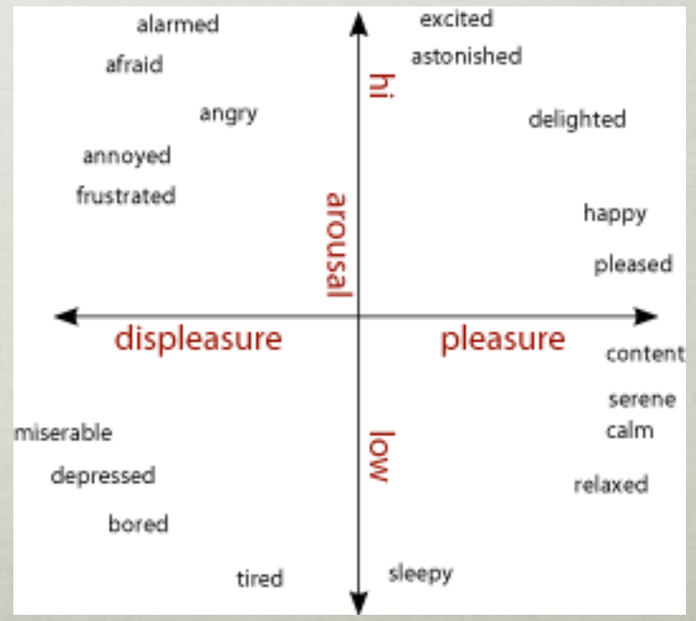
What are Expressions?

- Facial Displays are faces with a different appearance than neutral
- Facial Expressions are facial displays that express emotion
- Facial Displays can also express other messages, e.g. social signals

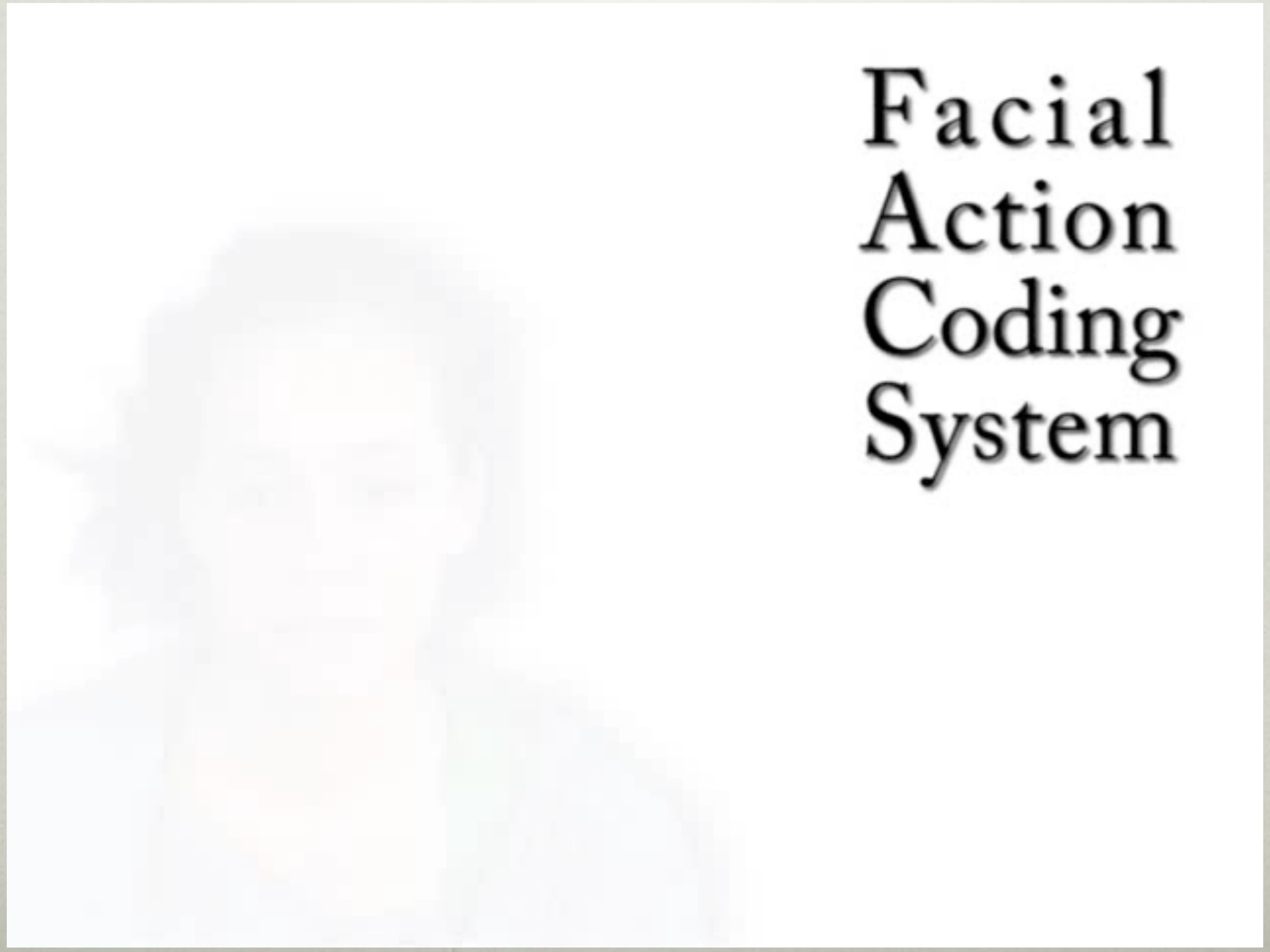


Emotion Theory

- There are a number of different theories on how humans perceive and feel emotions
- The most commonly used models are the six basic emotions, and Russell's Valence/Arousal/Dominance (VAD) model
- Both are message judgment models and inherently subjective
- Appraisal theory is very interesting, but complex



FACS



AHBU Tasks

- Discrete Expressions of Emotion (e.g. 6 basic emotions)
- FACS Action Units: Facial muscle actions
- Dimensional Affect (Valence/Arousal/Dominance)
- Higher-order behaviour
    - Severity of depression
    - ADHD/ASD diagnosis
    - Engagement estimation
    - Fatigue estimation

State of the Art

FERA 2011

- AU Sub-Challenge: Frame by frame detection of 12 Action Units

  87 Training videos

  71 Test videos

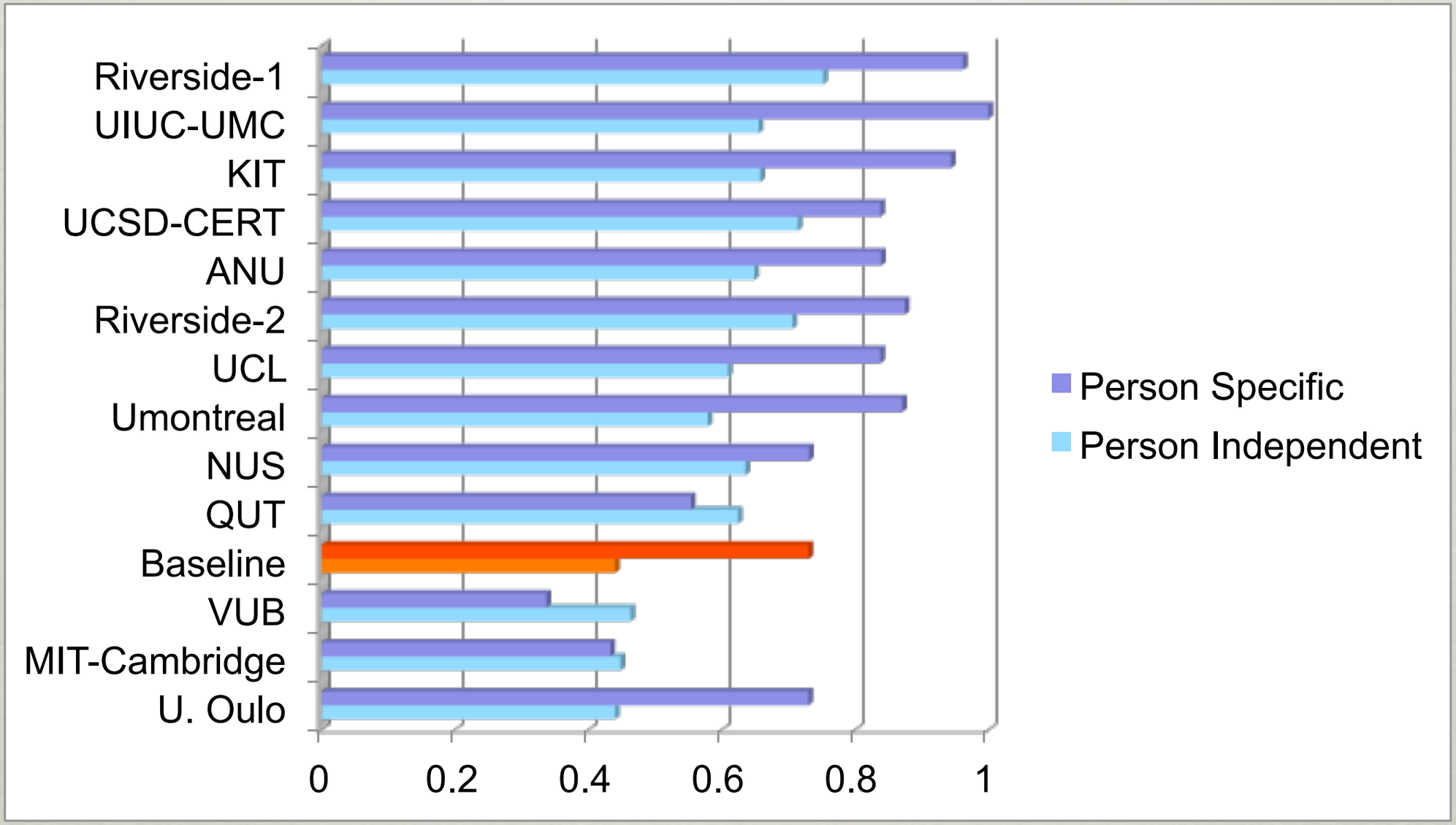
- Emotion Sub-Challenge: Event detection of 5 discrete emotions (Anger, Fear, Joy, Relief, Sadness)

  155 Training videos

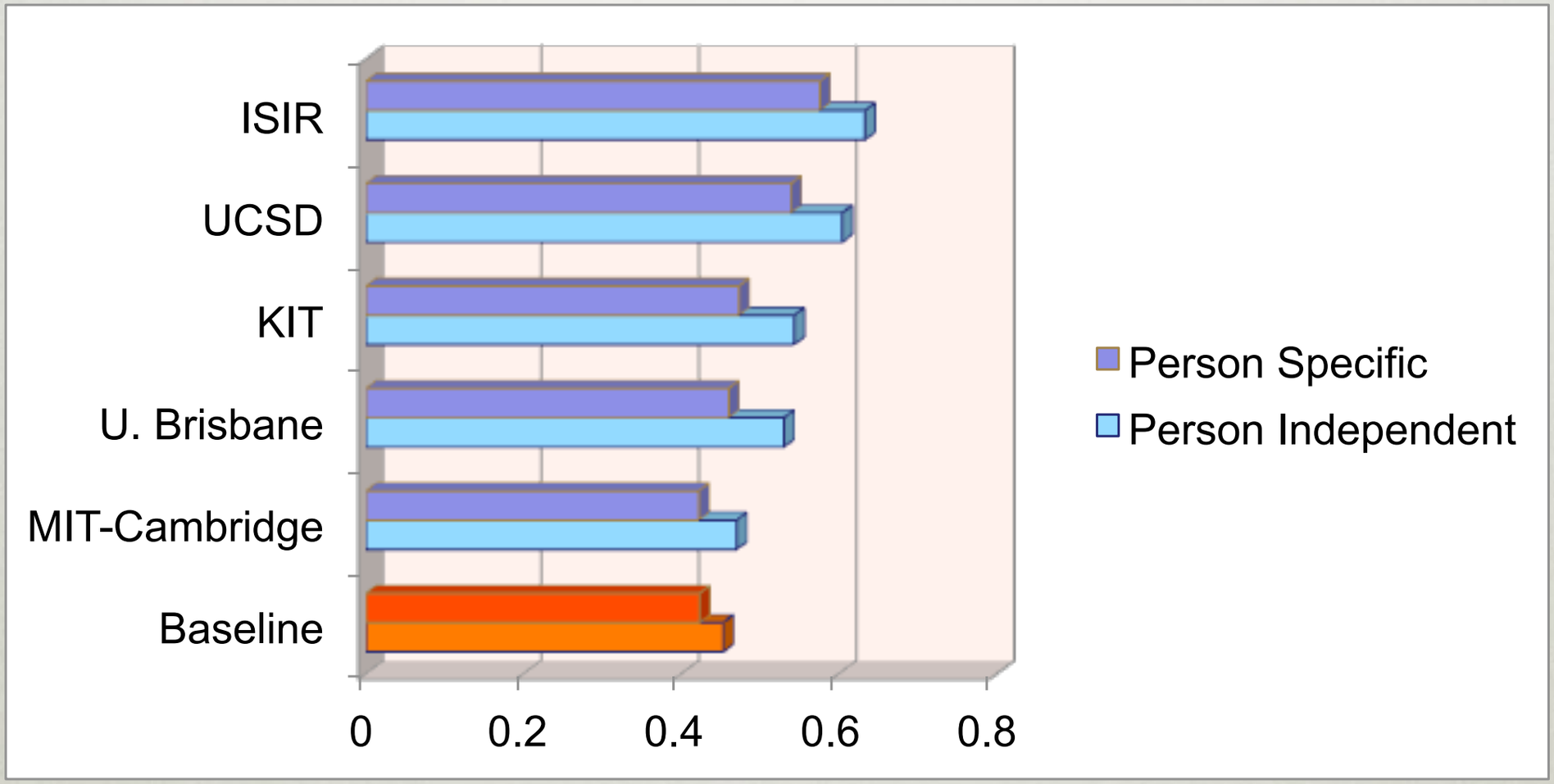
  134 Test videos

Michel Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer, 'Meta-Analysis of the First Facial Expression Recognition Challenge', IEEE Transactions on Systems, Man, and Cybernetics-B, vol. 42(4), pp. 966-979, 2012.

Discrete Emotions



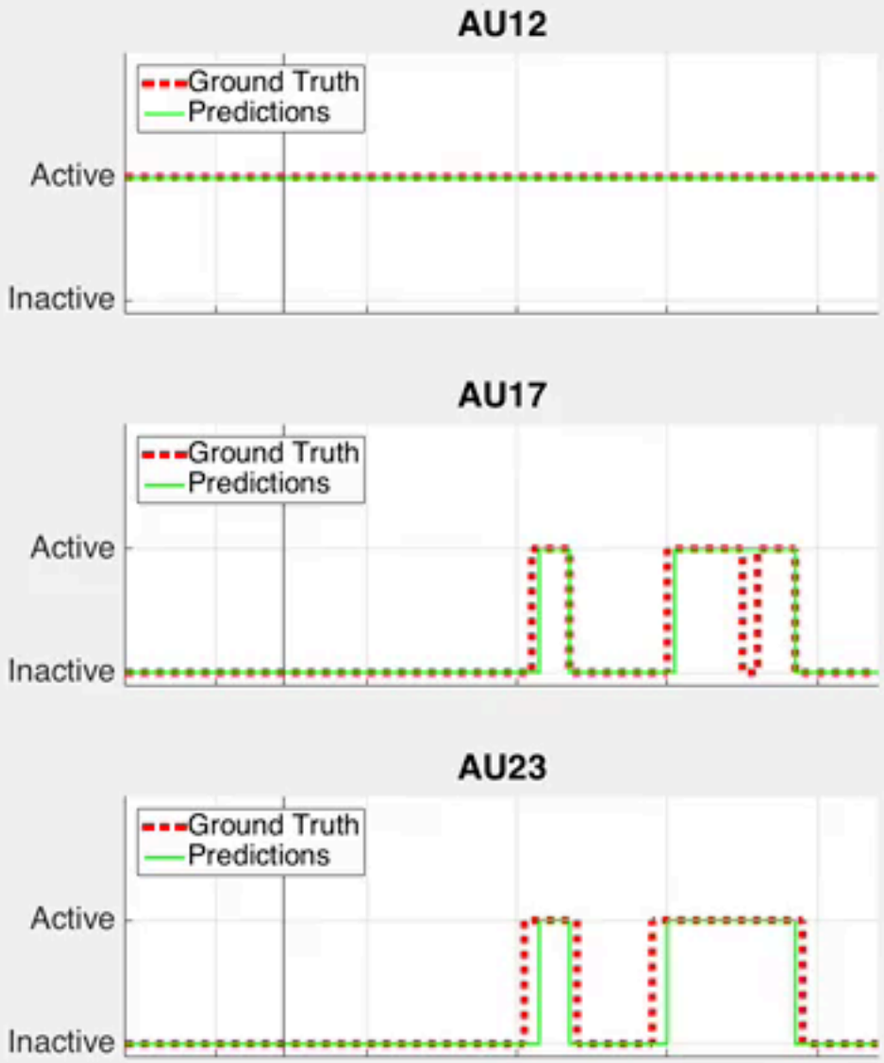
Action Units



FERA 2015: Occurrence

Occurrence sub-challenge results
(weighted F1 score over SEMAINE and BP4D)

AUs	EPFL-LTS5	Rainbow Group	VicarVision	Appearance baseline	Geometric baseline
Weighted Mean	0.499	0.481	0.465	0.400	0.444

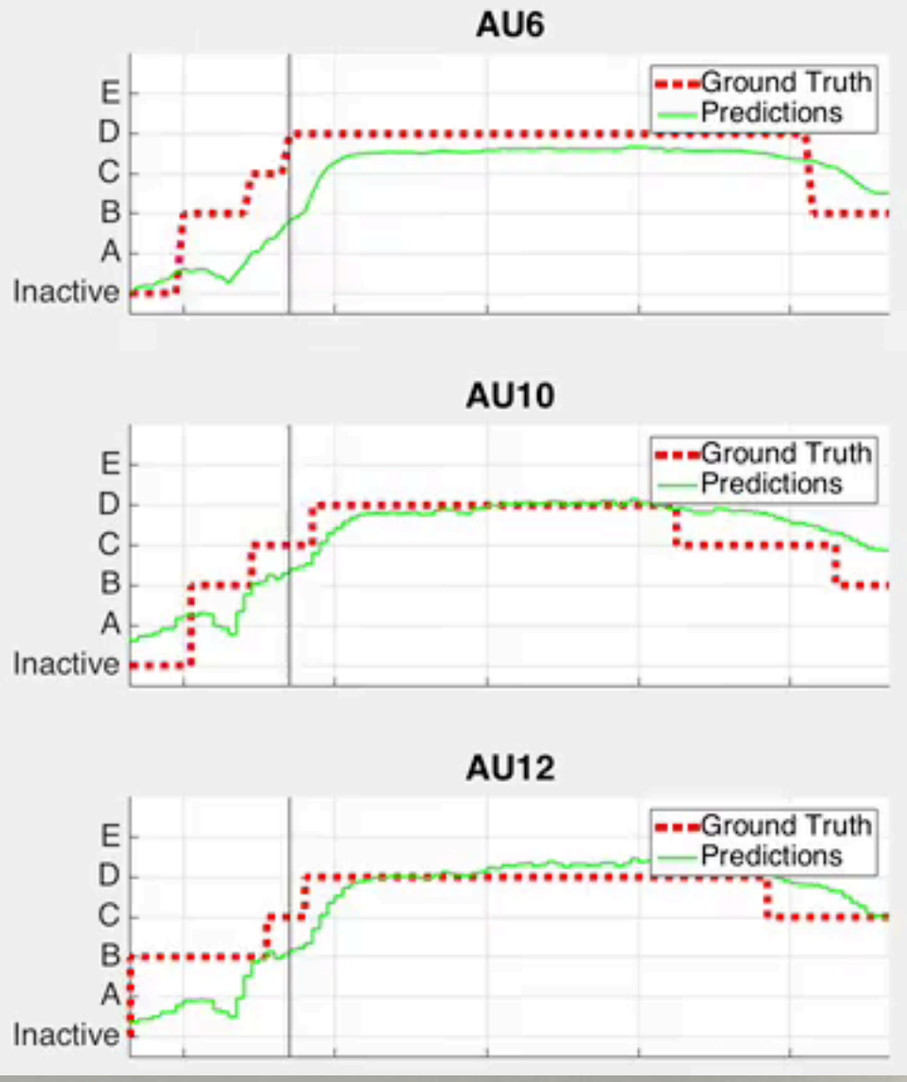

AU12
Ground Truth
Predictions
Active
Inactive
AU17
Ground Truth
Predictions
Active
Inactive
AU23
Ground Truth
Predictions
Active
Inactive

EPFL-LTS5

FERA 2015: Intensity

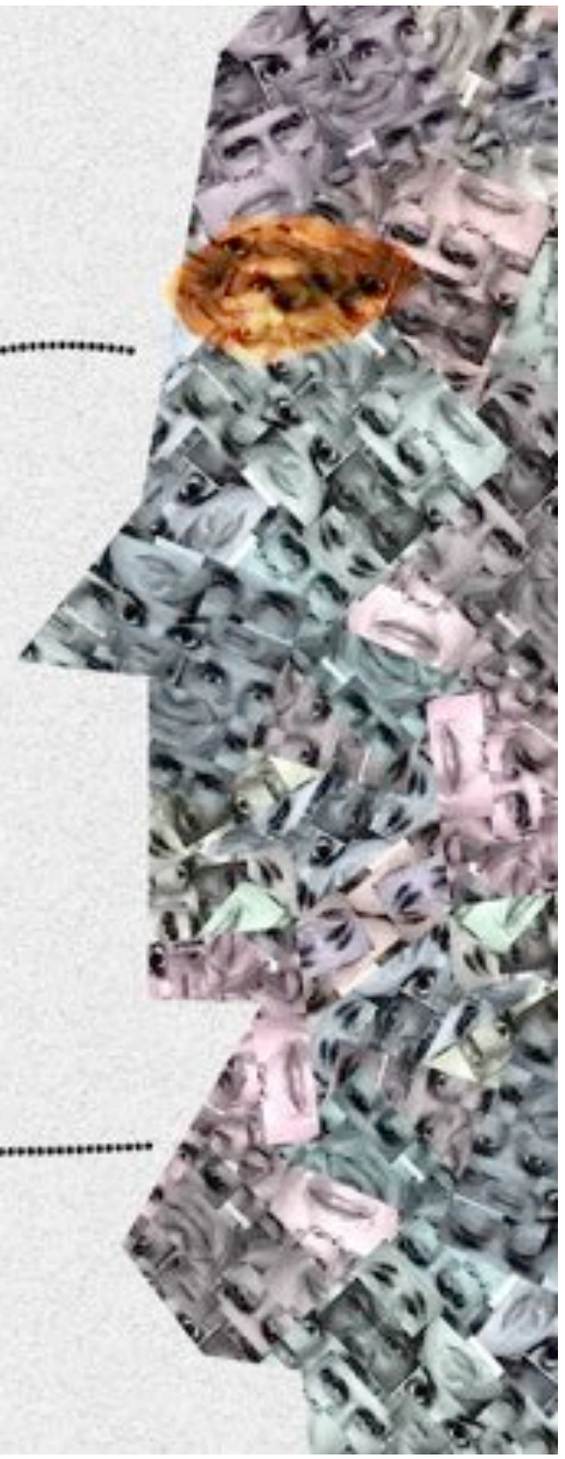
Fully automatic intensity sub-challenge results (ICC score, BP4D database)

AUs	ISIR	Rainbow Group	KIT	VicarVision	LaBRI	Appearance baseline	Geometric baseline
AU6	0.787	0.719	0.678	0.664	0.720	0.622	0.670
AU10	0.802	0.718	0.731	0.734	0.720	0.656	0.732
AU12	0.861	0.828	0.826	0.788	0.784	0.767	0.780
AU14	0.711	0.546	0.533	0.549	0.277	0.389	0.586
AU17	0.443	0.377	0.308	0.329	0.268	0.168	0.144
Mean	0.721	0.638	0.615	0.613	0.554	0.520	0.582


AU6
E
D
C
B
A
Inactive
Ground Truth
Predictions
AU10
E
D
C
B
A
Inactive
Ground Truth
Predictions
AU12
E
D
C
B
A
Inactive
Ground Truth
Predictions

FERA 2017

Facial Expression Recognition and Analysis challenge



AU OCCURRENCE

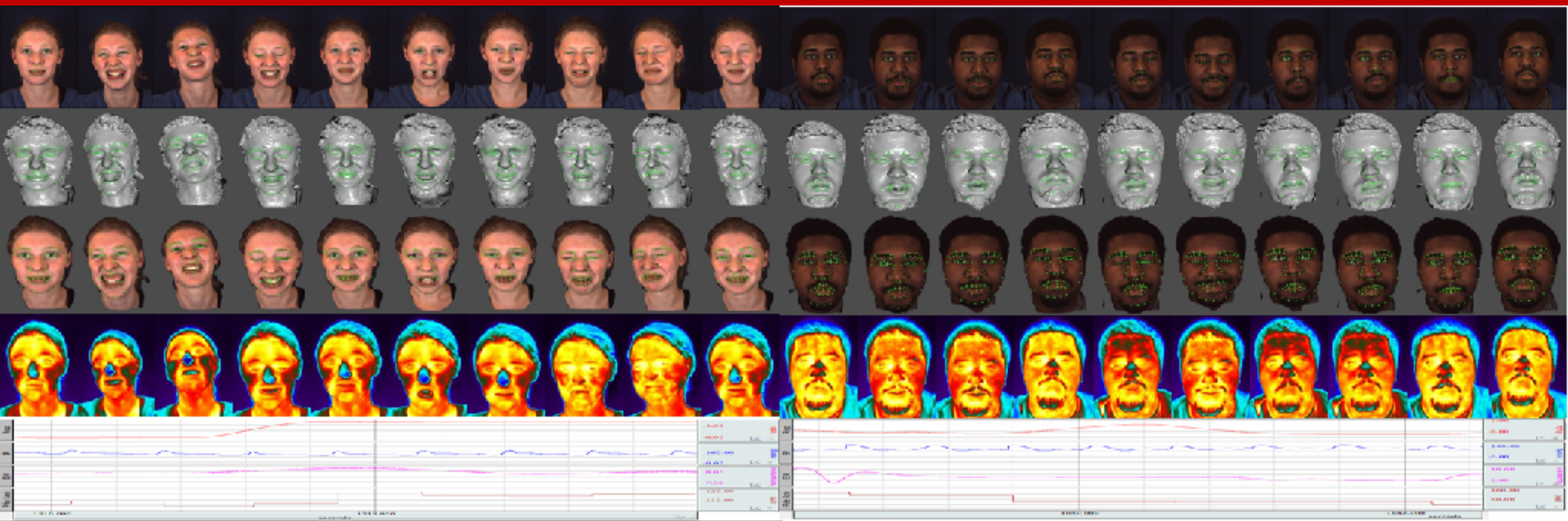
AU INTENSITY ESTIMATION

@FG2017, Washington, USA

Motivation

- Most Facial Expression Recognition systems focus on the binary occurrence of expressions from frontal-view images
- Yet in real-scenario data, head pose can vary wildly and low-intensity expressions are common
- Two FACS AU based sub-challenges are addressed:
  - **Occurrence Sub-Challenge:** Action Units have to be detected on frame basis in 9 different facial views (F1-Measure)
  - **Fully Automatic Intensity Sub-Challenge:** intensity of AUs must be predicted for each frame of a video in 9 different facial views (ICC score)

BP4D Dataset

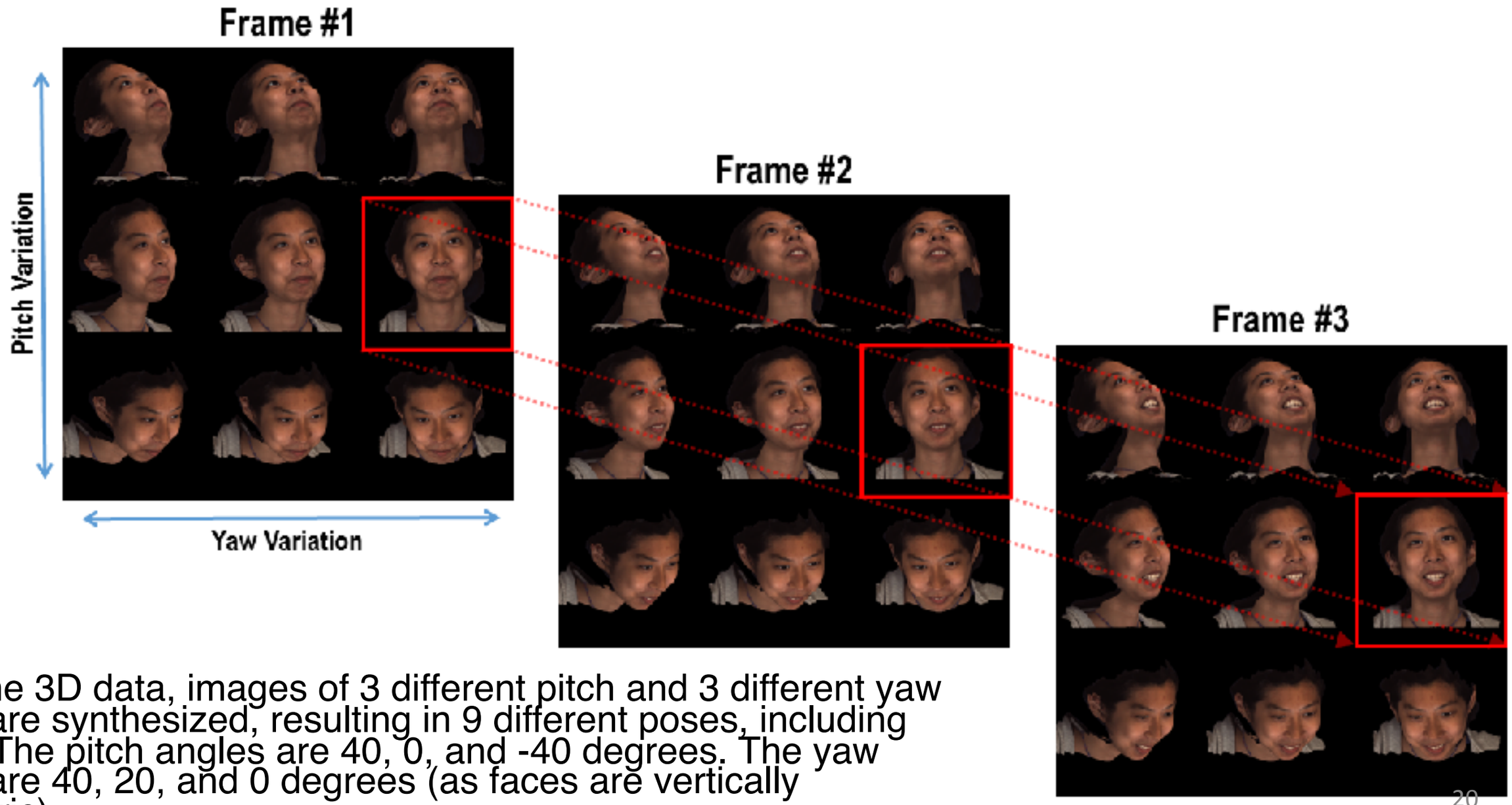


AU 12	AU 6+7+9+10+11+12+14+16+20	AU 1+6+7+9+10+11+12+14+23	AU 1+6+7+10+11+12+14+16+20+30	AU 1+6+7+10+11+12+14+15+23+24	AU 1+10+11+15+20	AU 1+6+7+10+12+14+23	N/A	N/A	AU 1+4+6+7+10+11+12+14+15+17+23

AU 1+2	AU 12+14+16	AU 12+15+17+23	N/A	AU 4+6+7+9+11+12	AU 6+7+10+14+12+14	AU 15+17+23+24+28+30+38	AU 11+12+28+38	AU 14+17+23+24+28	N/A

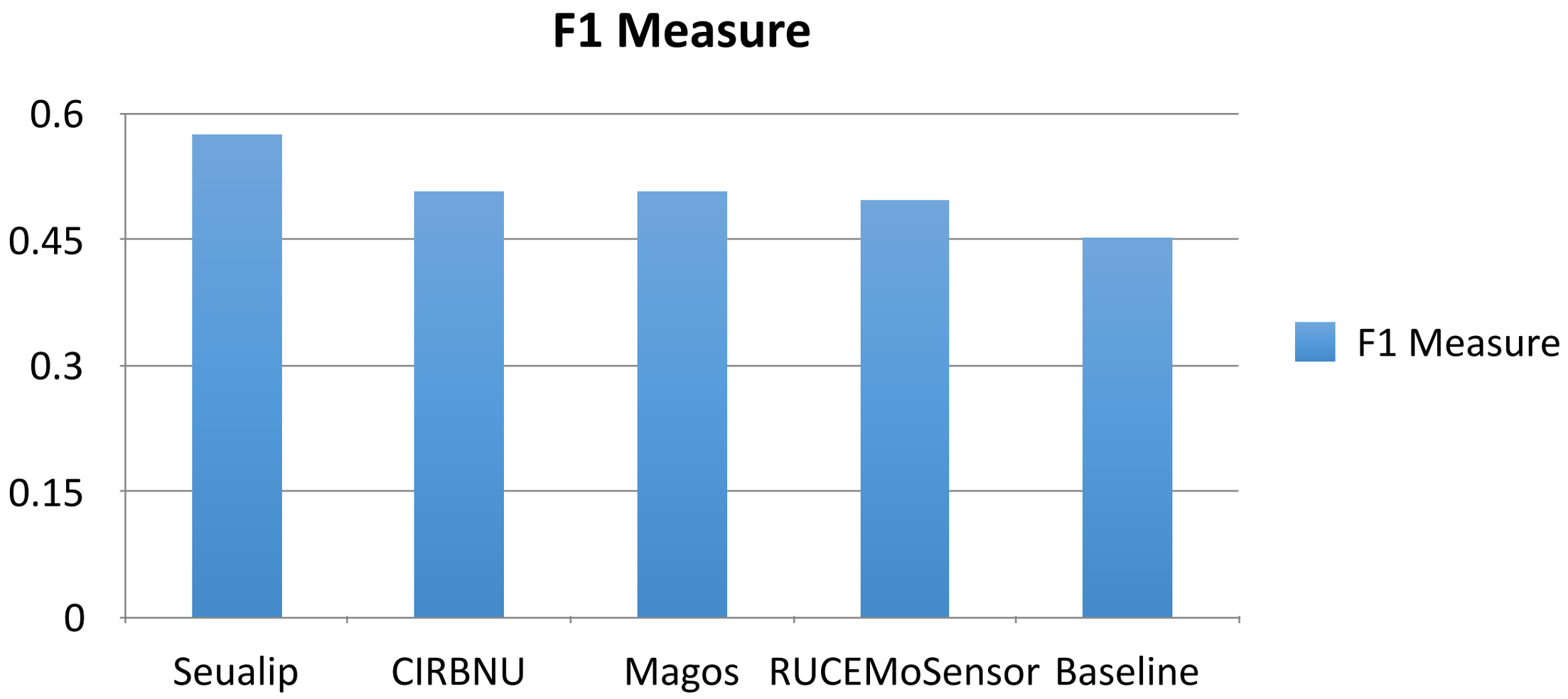
- The challenge data consists of data from 91 participants of the BP4D and BP4D+ datasets.
- Participants were recorded using a Di3D dynamic face capturing system during a series of eight emotion elicitation tasks, including happiness/amusement, sadness, surprise/startle, embarrassment, fear/nervous, physical pain, anger/upset, and disgust.

Data generation

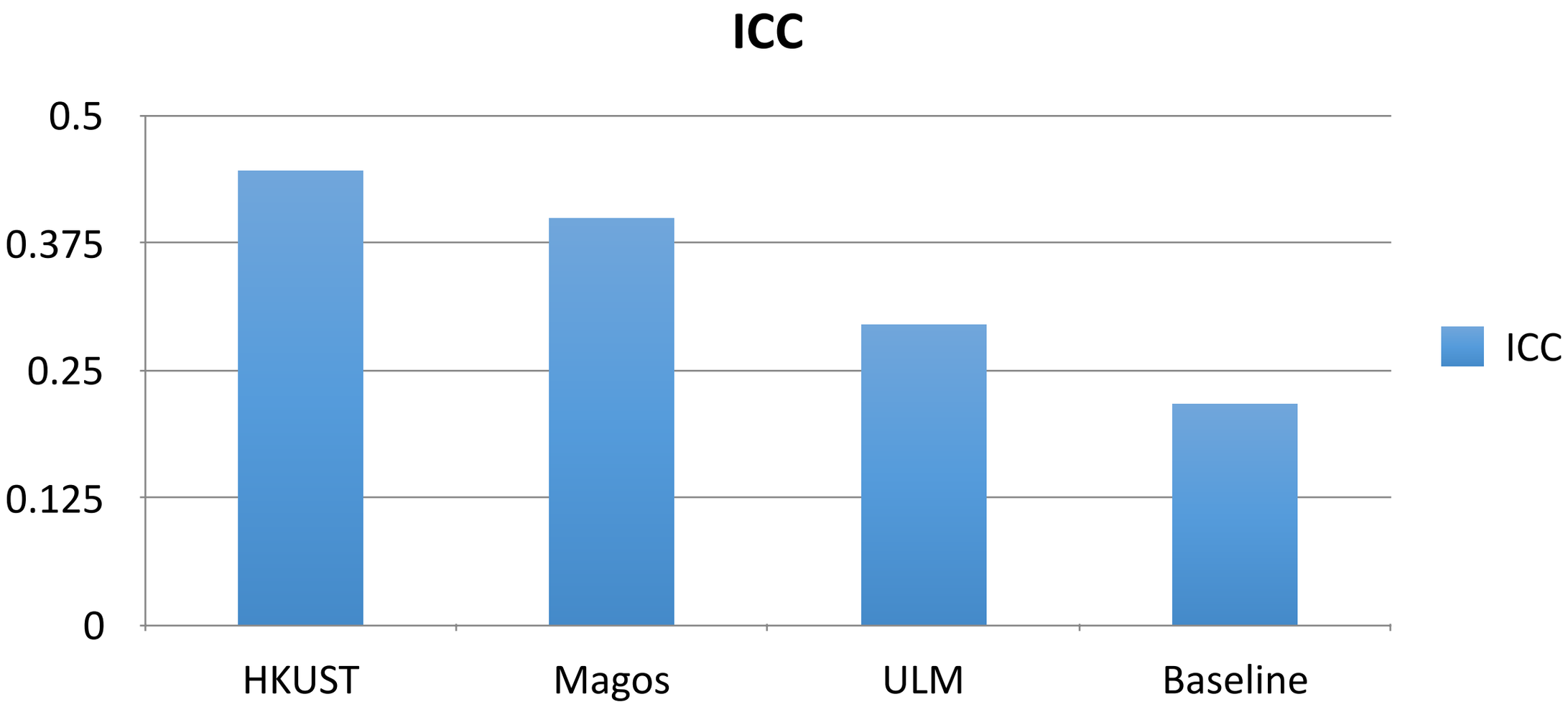


- Using the 3D data, images of 3 different pitch and 3 different yaw angles are synthesized, resulting in 9 different poses, including frontal. The pitch angles are 40, 0, and -40 degrees. The yaw angles are 40, 20, and 0 degrees (as faces are vertically symmetric).

Occurrence detection overview



Occurrence detection overview



Audio-Visual Dimensional Affect

AVEC 2012

4 affective dimensions: *Activation, Expectation, Power, Valence*

2-8 labellers for all sessions

68.4% > 5 raters 82% > 2 raters

Fully Continuous Sub-Challenge (FCSC):

Continuous time (prediction for every video frame)

Continuous in value (real number, range -1 to +1, expectancy from 0 to 100)

Björn Schuller, Michel F. Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic, 'AVEC 2012 – The Continuous Audio/Visual Emotion Challenge', proc. ACM Int'l Conf. Multimodal Interaction, pp. 449-456, October 2012

Audio-Visual Dimensional Affect

AVEC 2012

4 affective dimensions: *Activation, Expectation, Power, Valence*

2-8 labellers for all sessions

68.4% > 5 raters 82% > 2 raters

Fully Continuous Sub-Challenge (FCSC):

Continuous time (prediction for every video frame)

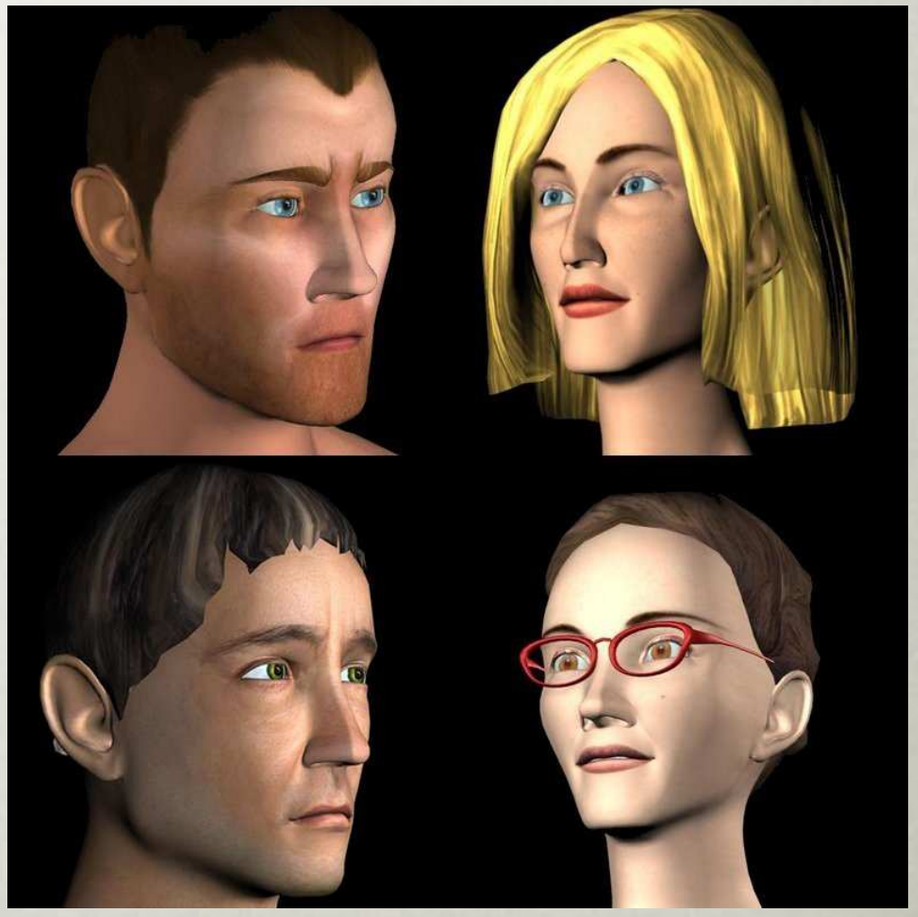
Continuous in value (real number, range -1 to +1, expectancy from 0 to 100)

Björn Schuller, Michel F. Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic, 'AVEC 2012 – The Continuous Audio/Visual Emotion Challenge', proc. ACM Int'l Conf. Multimodal Interaction, pp. 449-456, October 2012

SEMAINE Database

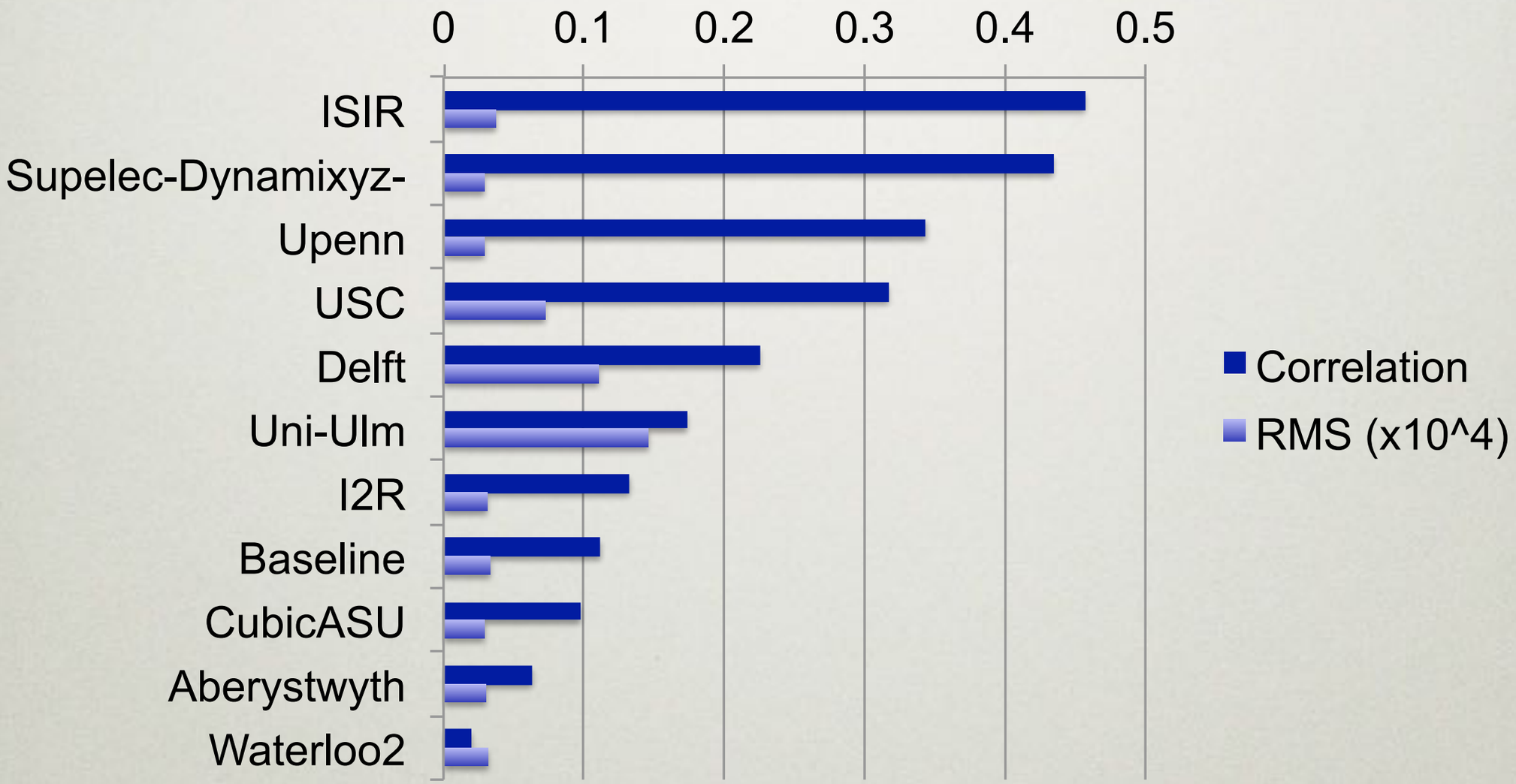
High-quality recordings of dyadic interactions between user and emotionally stereotyped operator

- 959 interactions
- 150 participants
- Database blocks:
  - Solid-SAL
  - Semi-automatic SAL
  - Automatic SAL

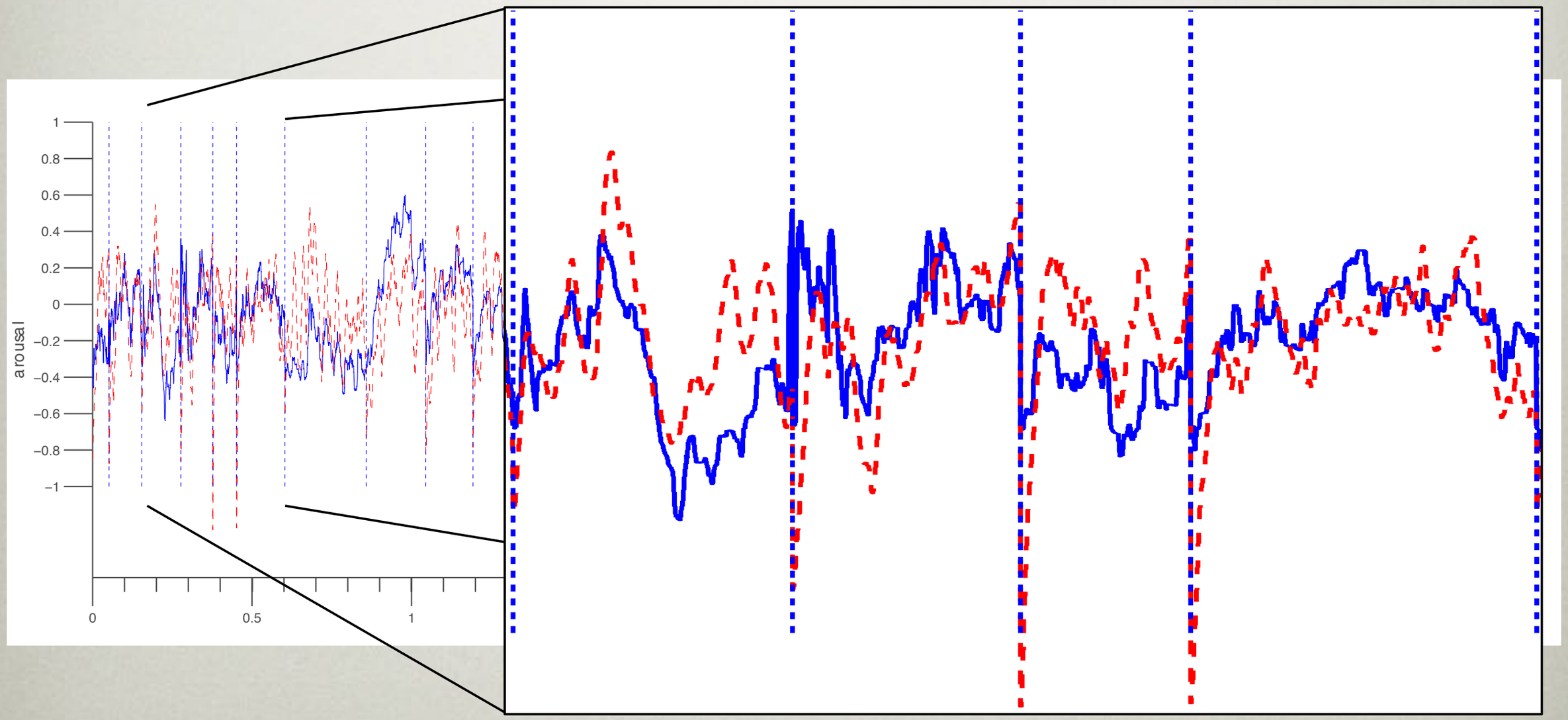


Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder, 'The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent', IEEE Transactions on Affective Computing, vol. 3, pp 5-17, 2012

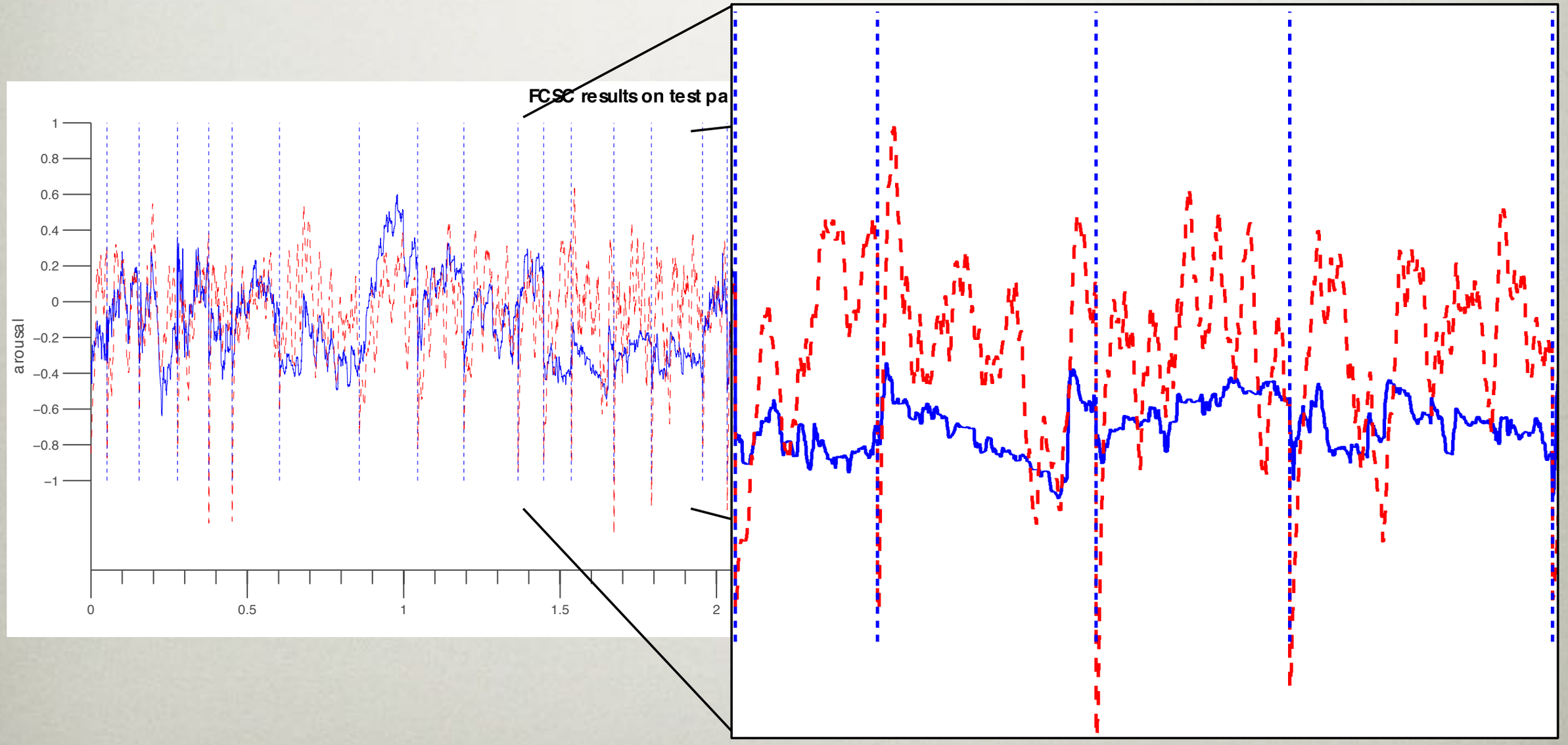
Fully Continuous Sub-Challenge



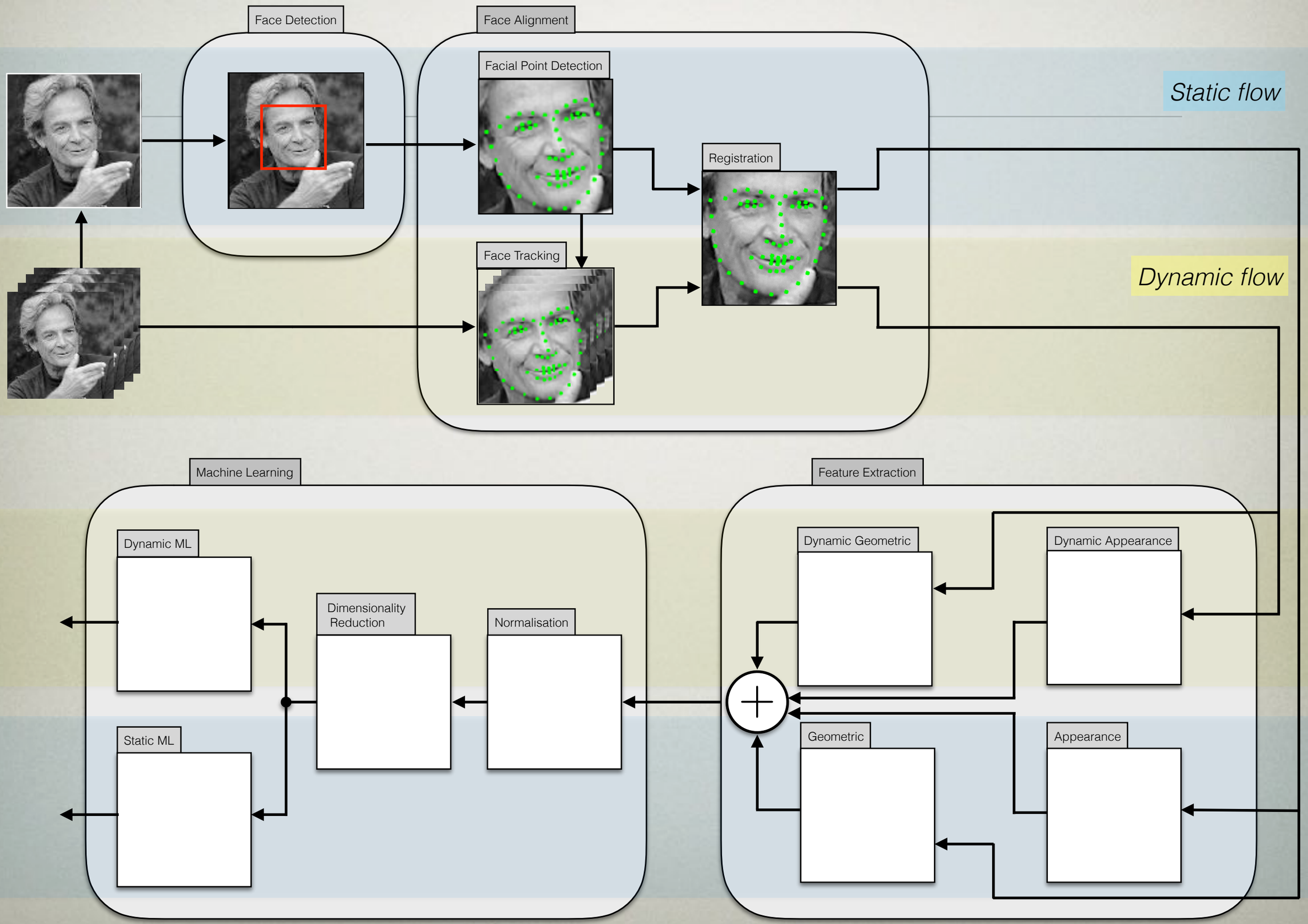
Visualisation of Best FCSC Result



And Some Poor Results



Classic Pipeline



Pre-Processing

- Goals:

    1. To detect location of the face
    2. To remove variation of head pose
        or to allow using mode-specific models
    3. To remove variations between people
    4. To remove variations within one person

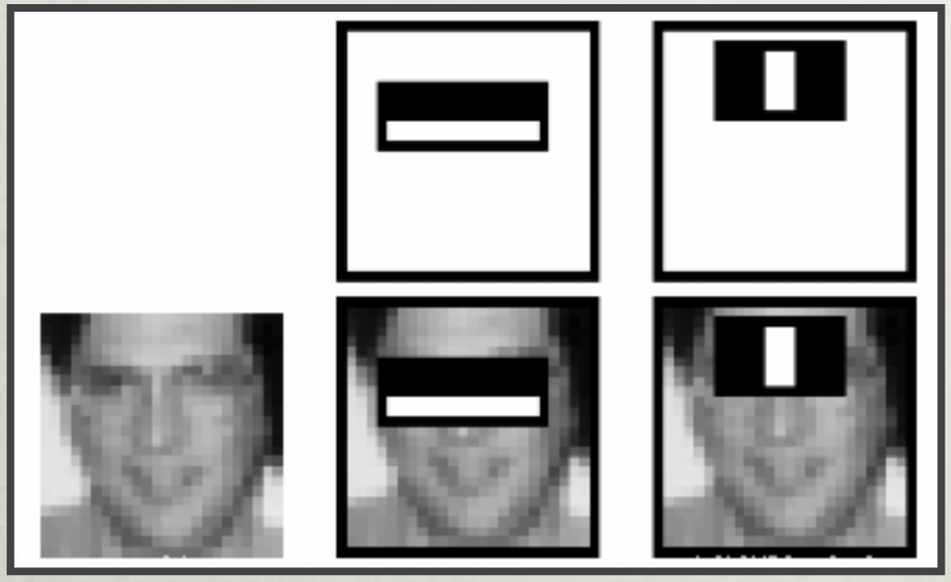
Face Detection

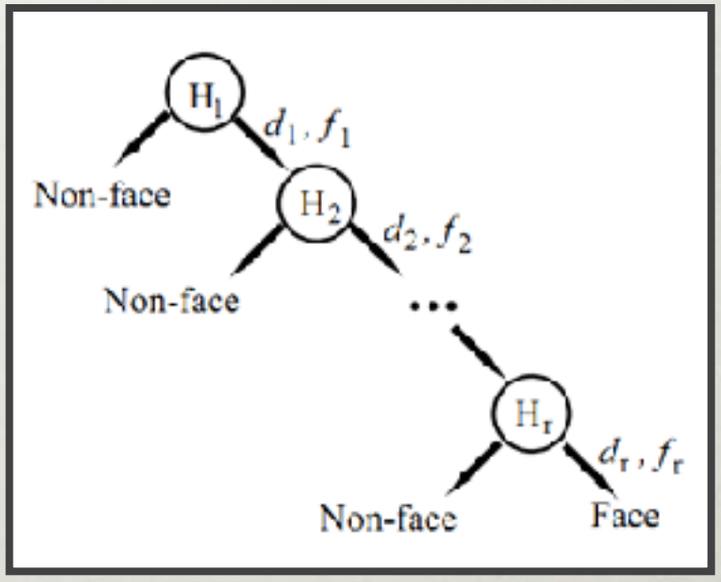
- Goals:
  1. To detect presence of any (and all) faces
  2. To detect location of the face (x,y location)
  3. To estimate scale of the face (height, width)

Viola & Jones

- Goals:

    1. To detect location of the face (x,y location)
    2. To detect extend of the face (height, width)





Viola & Jones (2003). *'Fast multi-view face detection'*, Technical report, Mitsubishi Electric Research Laboratory

Zhu & Rhamanan

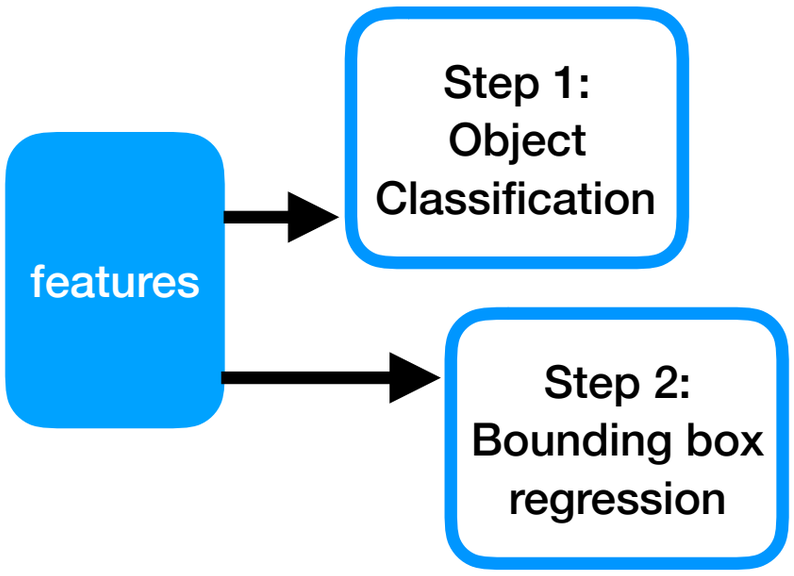
- Goals:

1. To detect location of the face (x,y location)

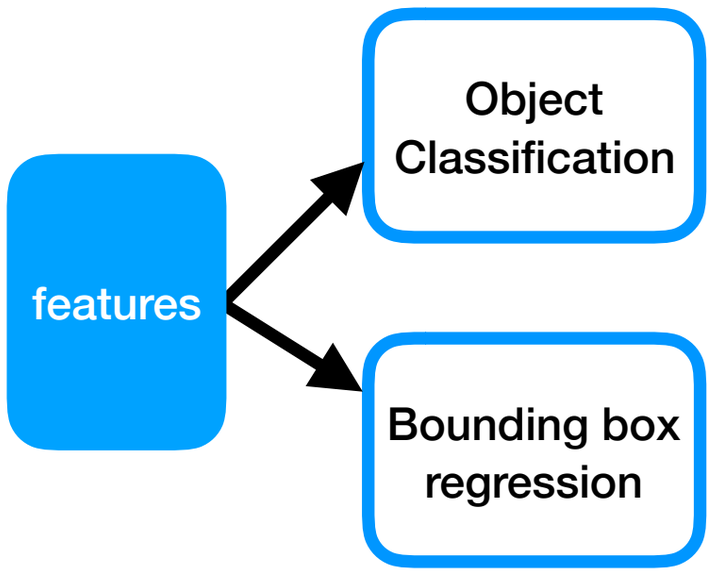
2. To detect extend of the face (height, width)

Zhu & Ramanan (2013). *'Face detection, pose estimation and landmark localization in the wild'* Computer Vision and Pattern Recognition 2013

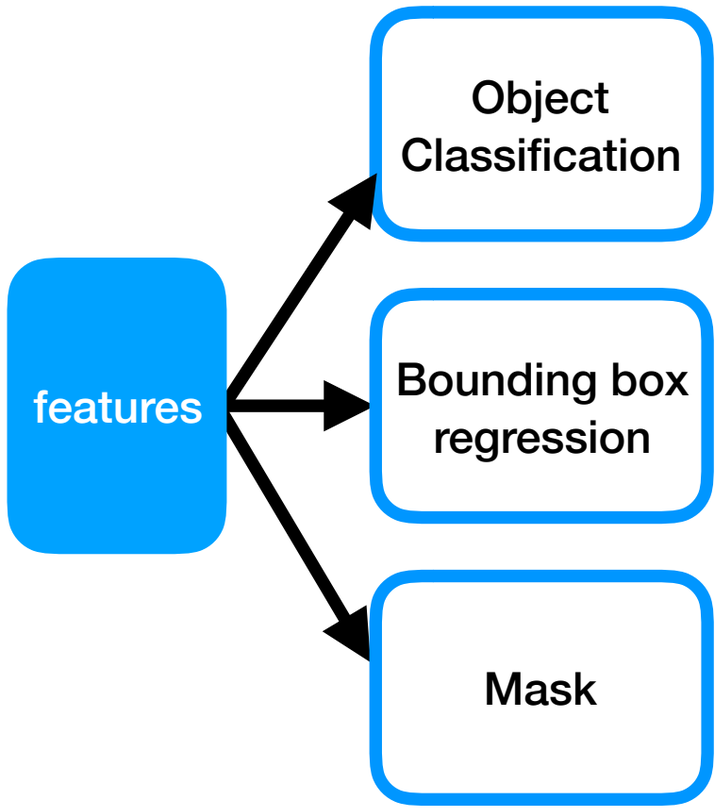
DEEP-LEARNED FACE DETECTION



(a) R-CNN

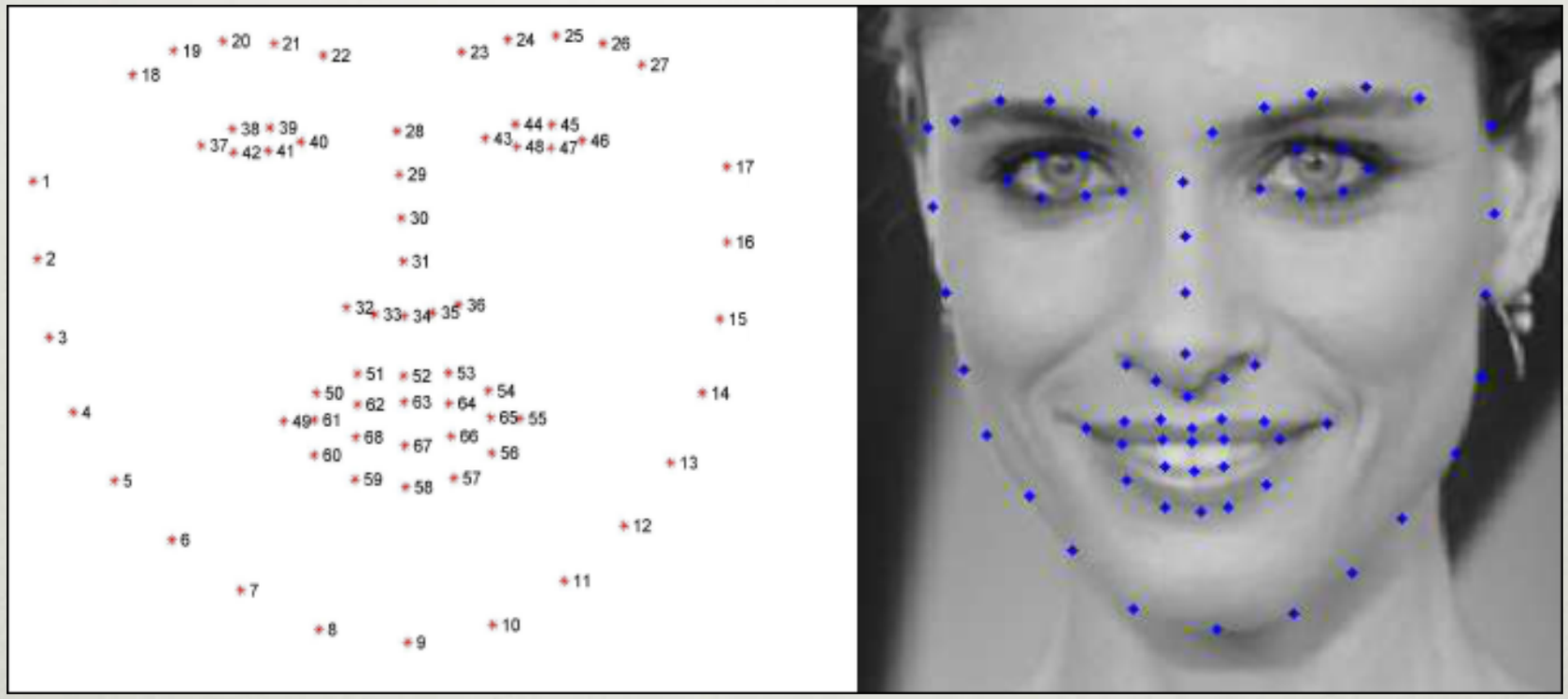


(b) fast(er) R-CNN



(c) mask R-CNN

Face Alignment

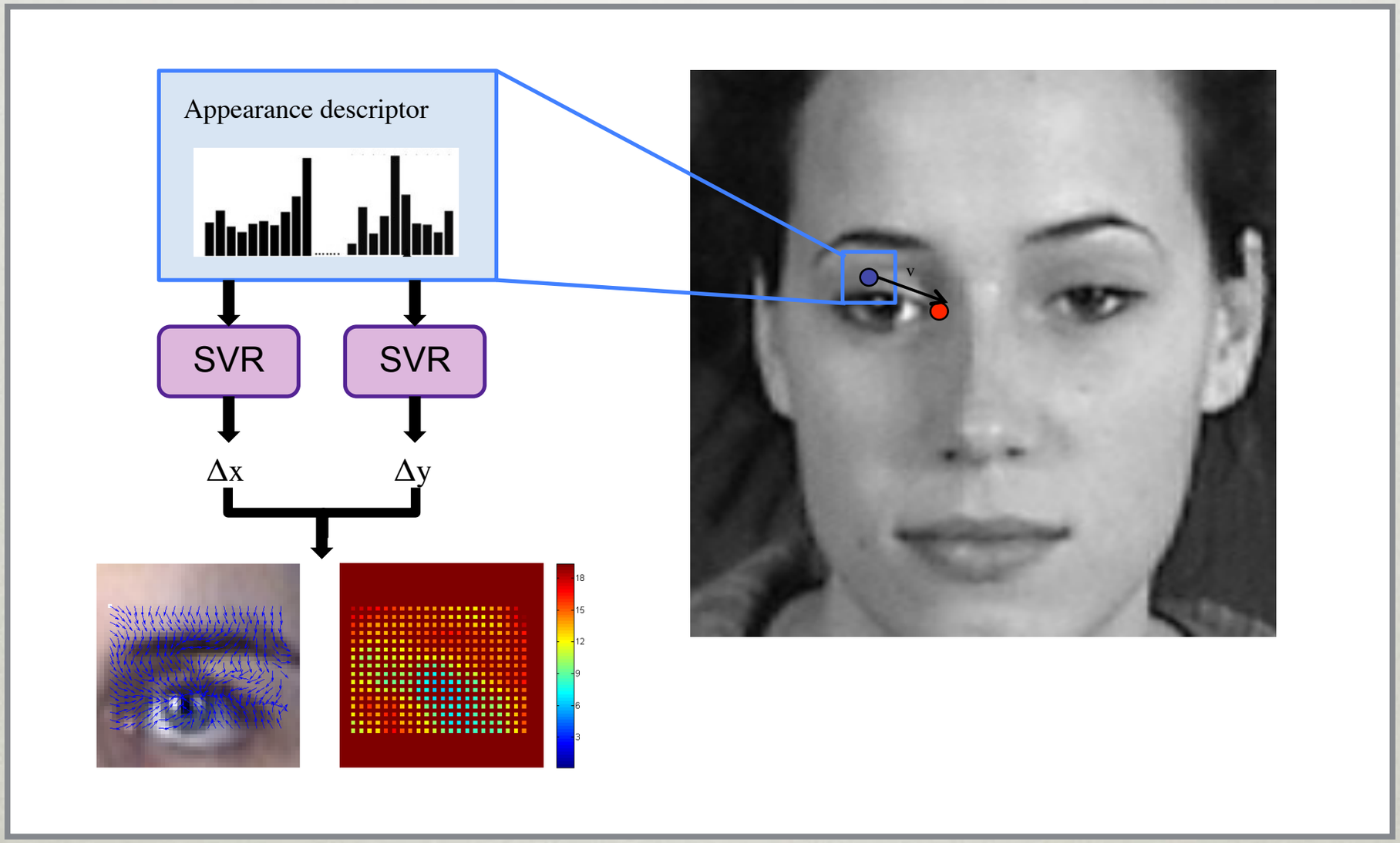


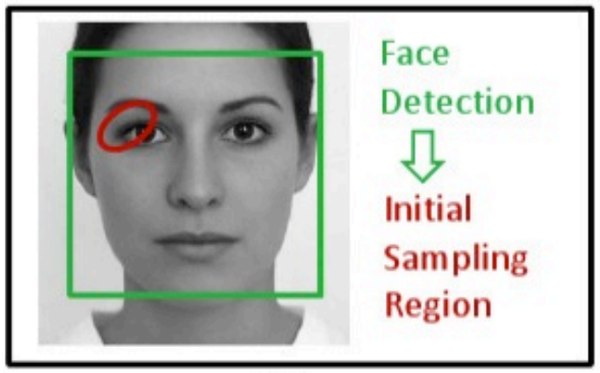
High variation in face images caused due to

- Different head poses
- Different facial expressions

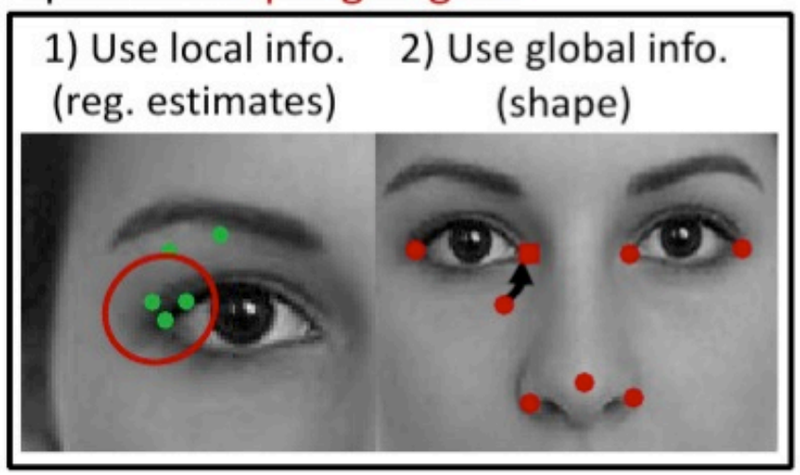
Results in non linear changes in the shape and appearance of facial points

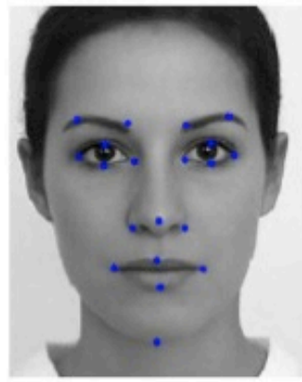
Regression-Based Approaches





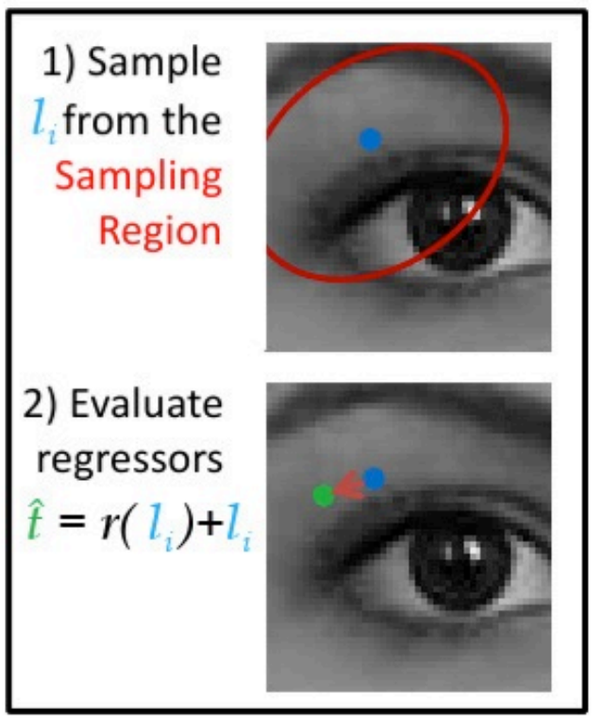
Update Sampling Region





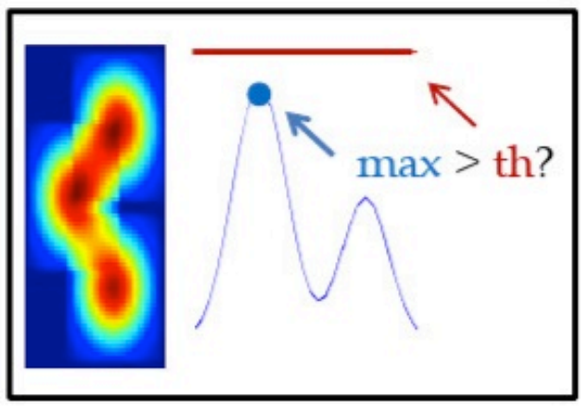
Early stop condition

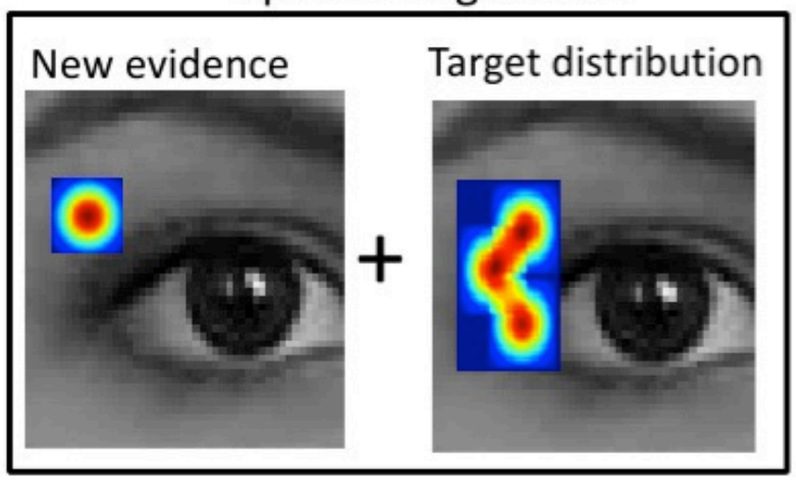
Obtain new estimate



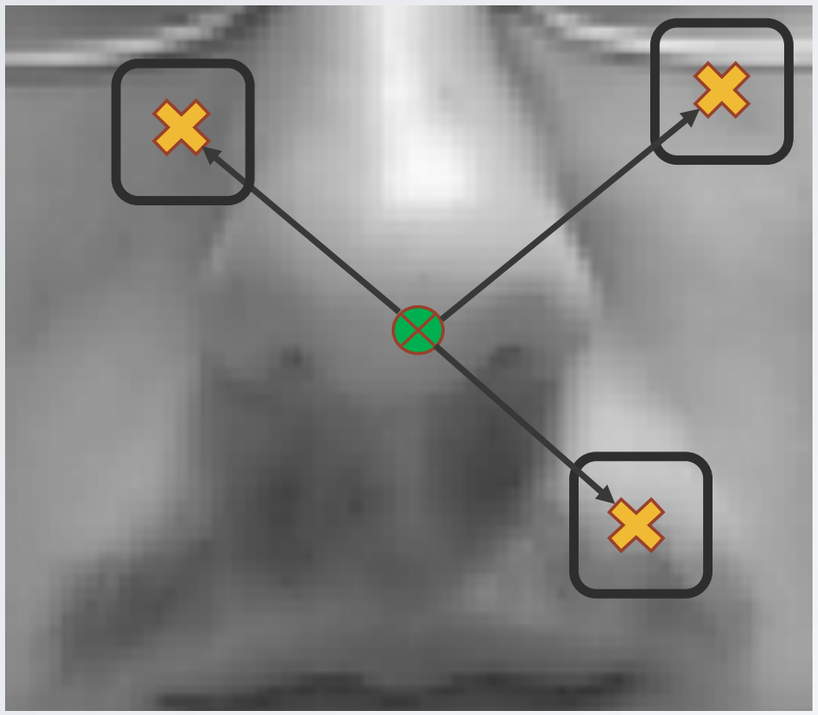
While *it* < *itmax*

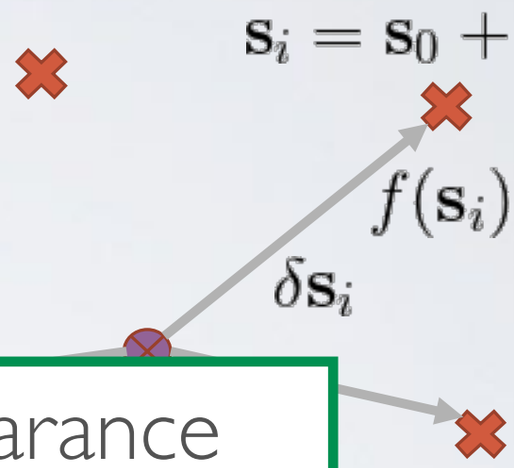
Update Target Dist.





LINEAR REGRESSION

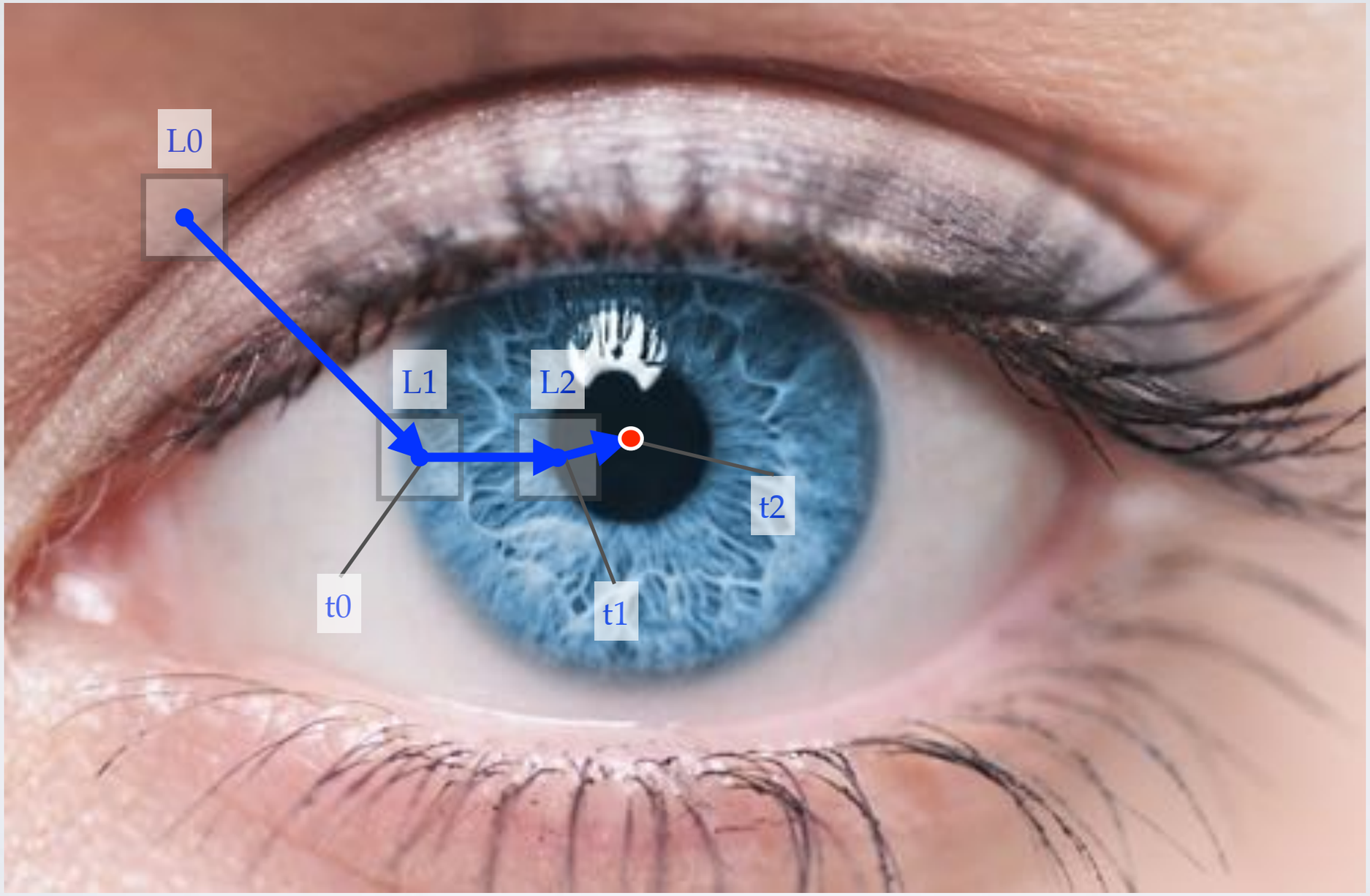




Some appearance descriptor (e.g. HOG)

$$\mathbf{R} = \arg\min_{\mathbf{R}} \sum_{k=1}^{K} \|\delta\mathbf{s}_k - \mathbf{R}f(\mathbf{s}_k)\|_2^2$$

CASCADED REGRESSION



INTRINSIC PROBLEMS

- Linear Regression generates biased models
- Supervised Descent Method (i.e. cascaded linear regression) is expensive to train
- Incremental learning is too slow

We propose incrementally learned cascaded continuous regression (iCCR) as a solution

E. Sánchez Lozano, G. Tzimiropoulos, B. Martinez, F. De la Torre and M.F. Valstar (2017). *"A Functional Regression Approach to Facial Landmark Tracking*, TPAMI [Code available!]

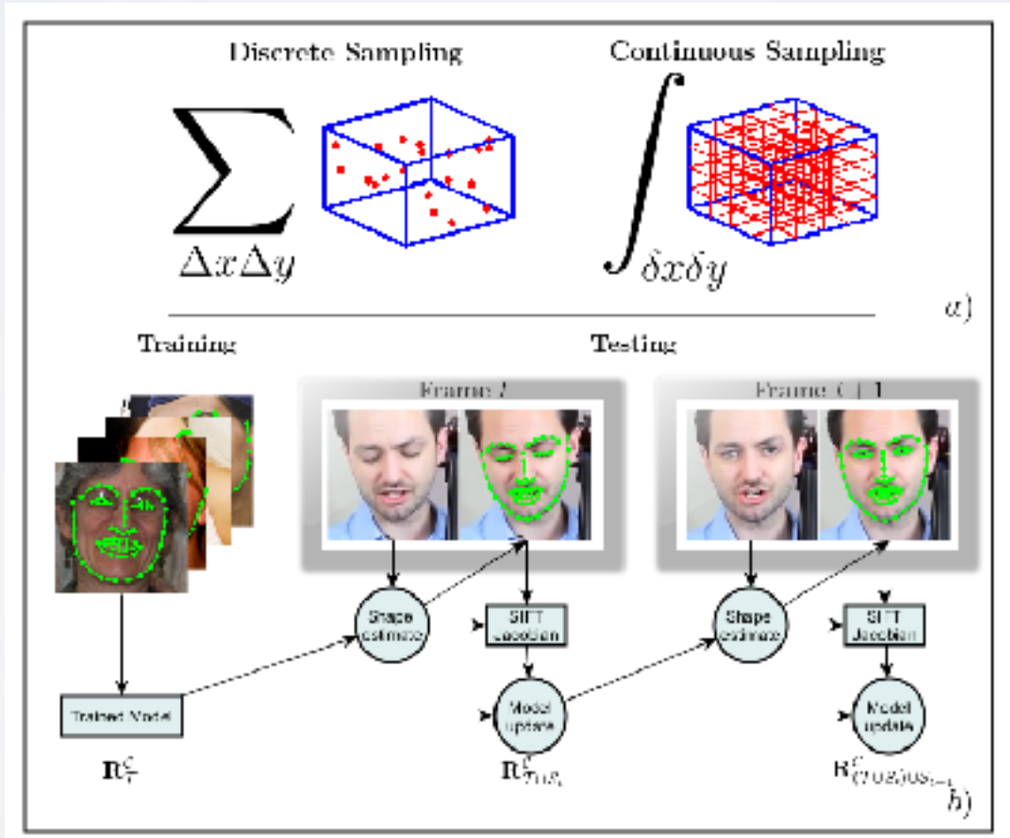
CONTINUOUS REGRESSION

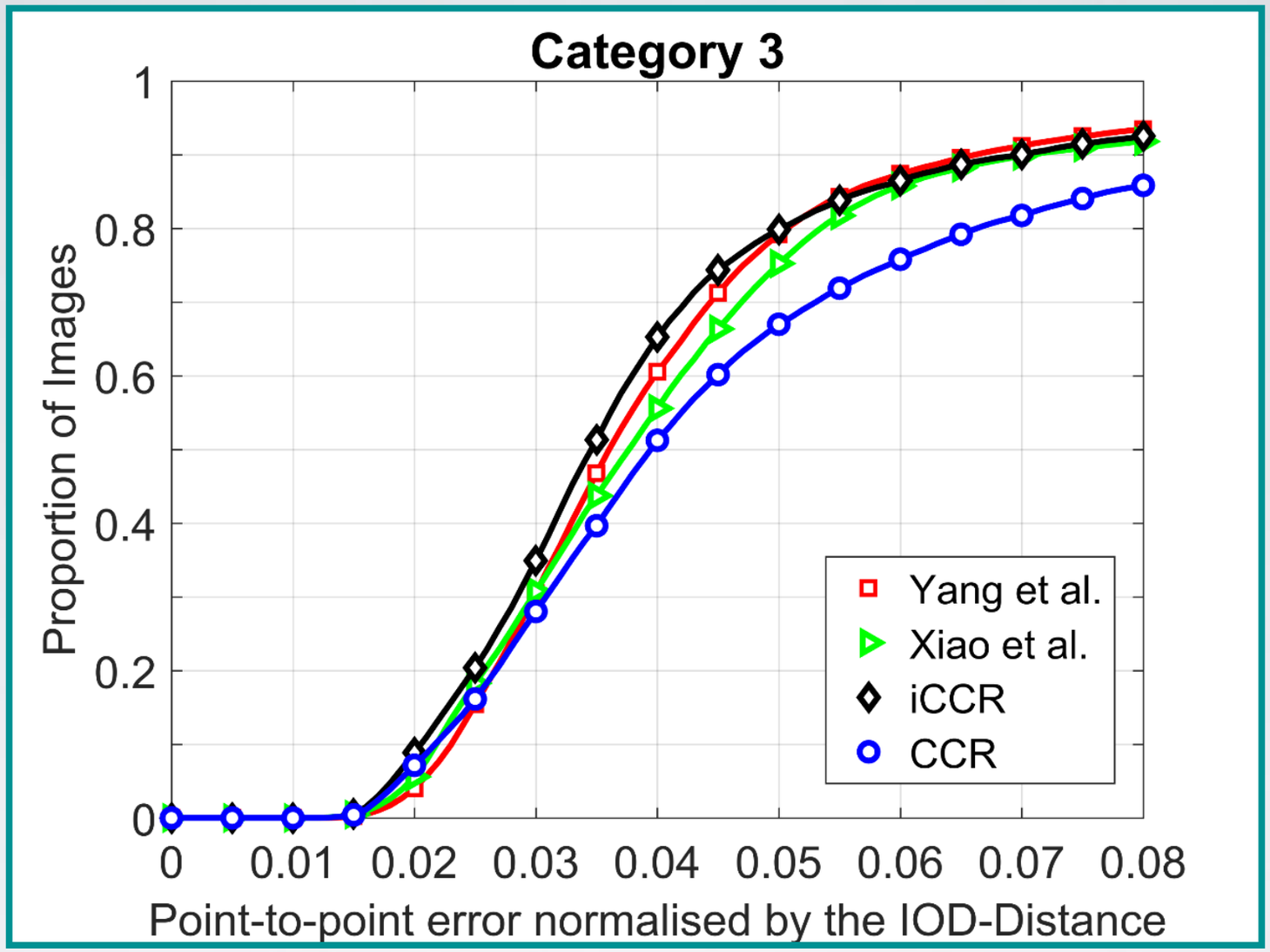
- We want to use *all* examples
- This is equivalent to treating them as part of a continuum
- Sum of samples becomes integral over space

$$\mathbf{R} = \arg\min_{\mathbf{R}} \sum_{k=1}^{K} \|\delta\mathbf{s}_k - \mathbf{R}f(\mathbf{s}_k)\|_2^2 \quad \Longrightarrow \quad \mathbf{R} = \arg\min_{\mathbf{R}} \int \|\delta\mathbf{s} - \mathbf{R}f(\mathbf{s}_0 + \delta\mathbf{s})\|_2^2 d\delta\mathbf{s}$$

FEATURE APPROXIMATION

- We can't sample all points (too expensive)
- We need to approximate input space with some bases
- Use first-order Taylor expansion: $f(\boldsymbol{s}_0 + \delta\boldsymbol{s}) \approx f(\boldsymbol{s}_0) + f'(\boldsymbol{s}_0)\delta\boldsymbol{s}$





- Small accuracy improvement over the state of the art, but computational complexity reduced by an order of magnitude
- Offline training time sped up by a factor 25
- Incremental learning time sped up by a factor 10

E. Sánchez Lozano, G. Tzimiropoulos, B. Martinez, F. De la Torre and M.F. Valstar (2017). *"A Functional Regression Approach to Facial Landmark Tracking*, TPAMI [Code available!]

EXAMPLE

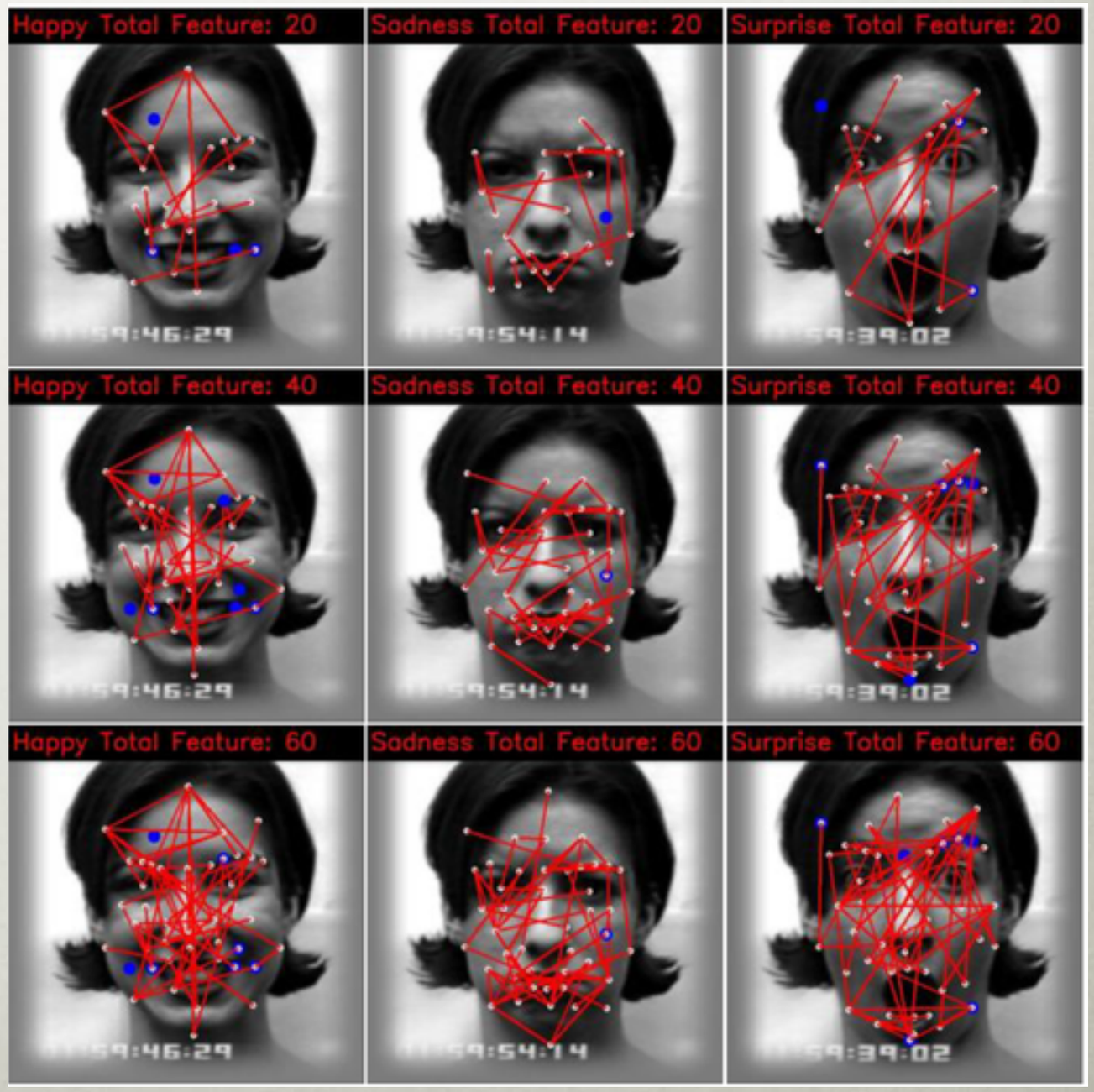
Video taken from ComputerPhile: Smile Detection using Local Binary Patterns

Computational Face Descriptors

- Types of features:
    - Learned:
        1. Convolutional Neural Networks
        2. Networks with memory: LSTM/GRU
    - Hand-crafted:
        1. Geometric Features
        2. Appearance Descriptors
        3. Hybrid

Geometric Features (Hand-Crafted)

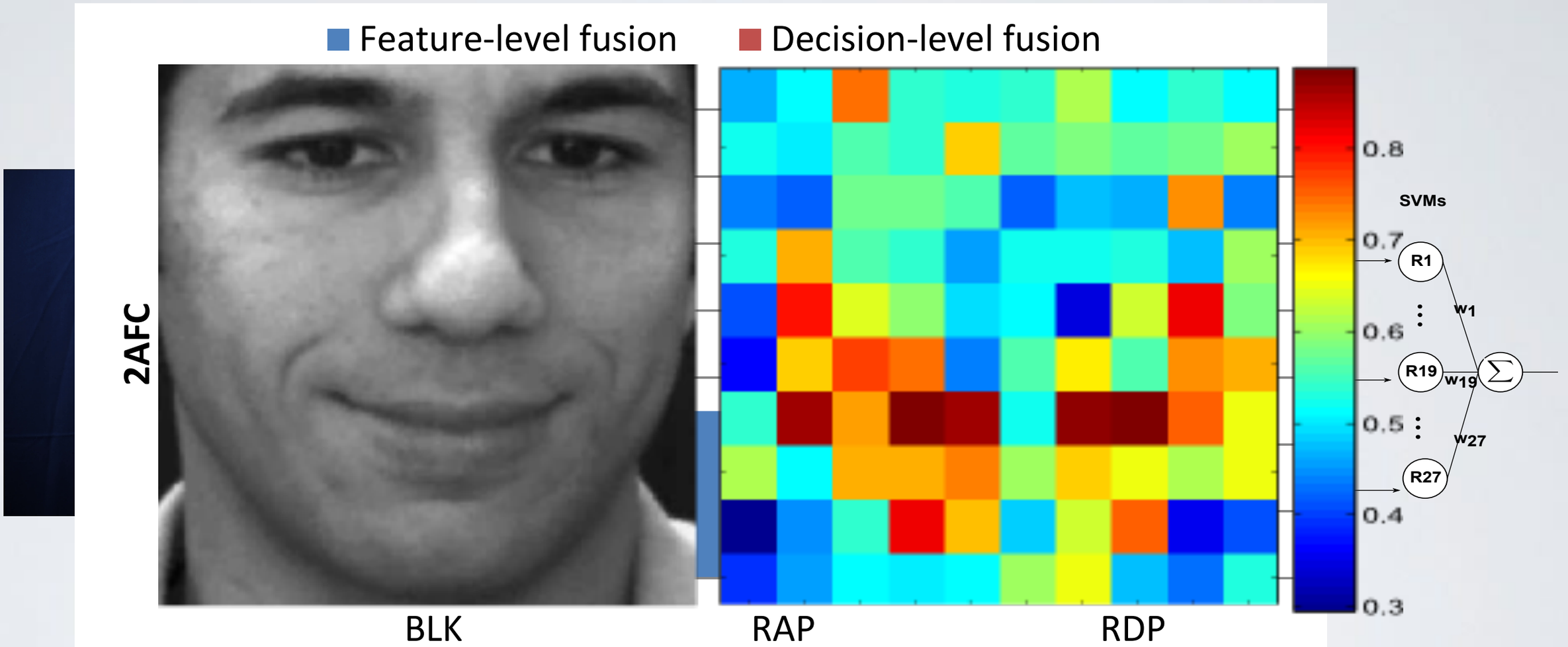
- Expert rules
- Simple rules
- Temporal extensions
- Features of pairs / triplets of points



Appearance Descriptors (Hand-Crafted)

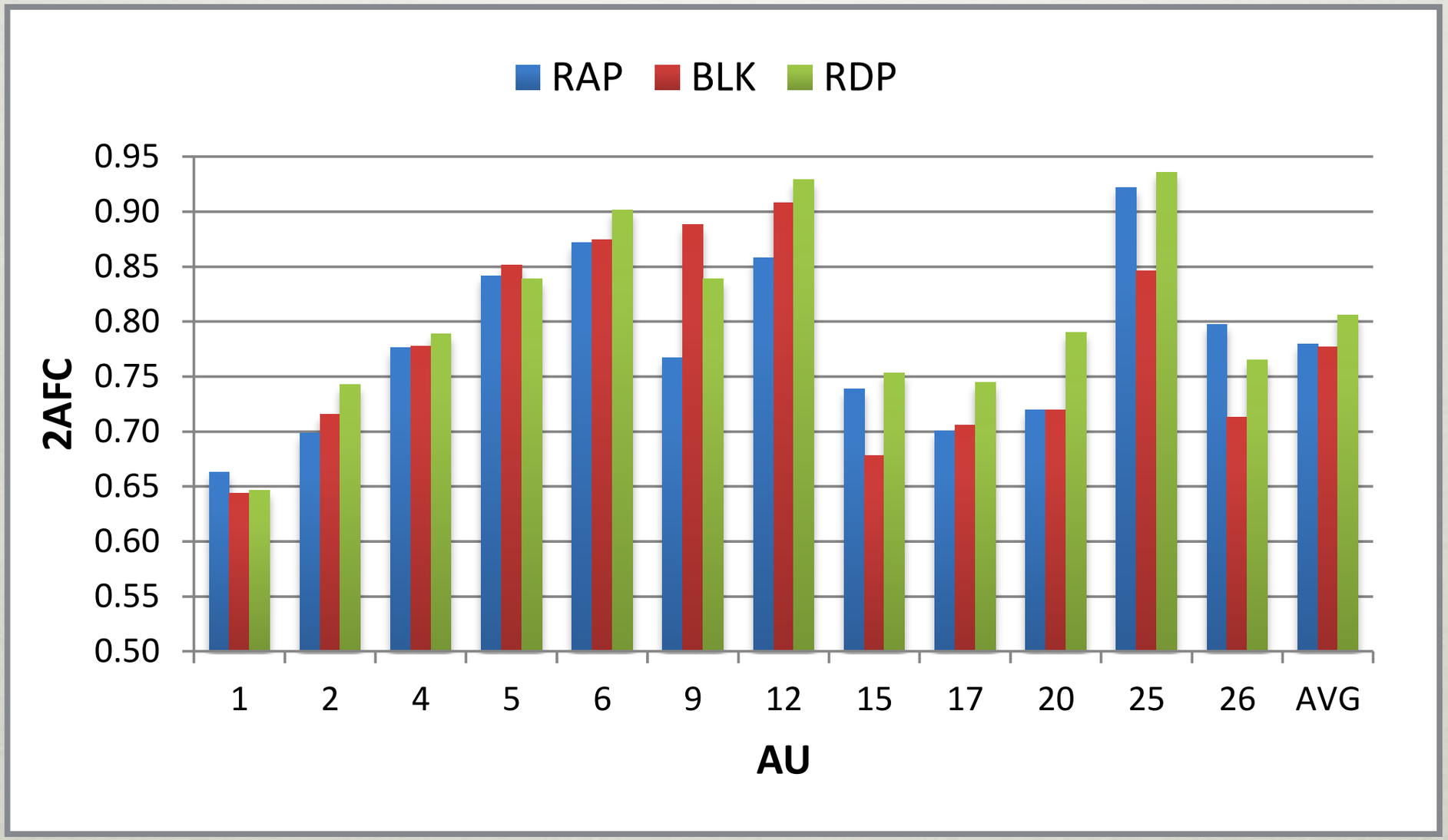
- Commonly used appearance descriptors:
  1. Gabor
  2. Local Binary Patterns
  3. HOG/SIFT

USE THE RIGHT REGION



Bihan Jiang, Brais Martinez, Michel Valstar (2014) 'Decision Level Fusion of Domain Specific Regions for Facial Action Recognition', ICPR

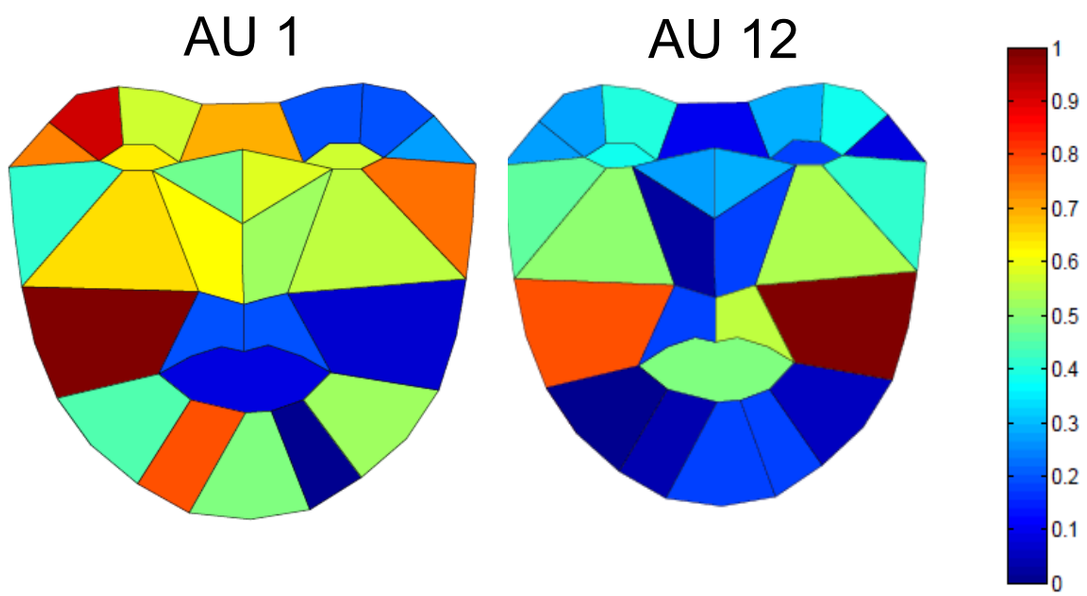
RDP Results



Jiang, Martinez, Valstar, and Pantic (2014), *'Decision Level Fusion of Domain Specific Regions for Facial Action Recognition'*, Proc. Int'l Conf. Pattern Recognition (ICPR)

ANNs vs SVM fusion

Importance of regions



SVM fusion vs ANN (DISFA)

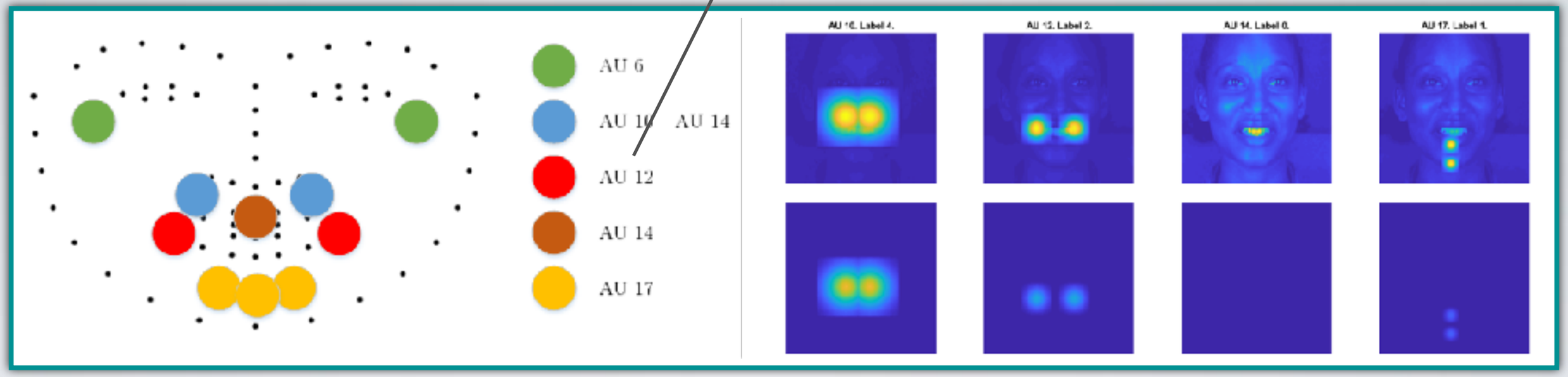
AU	Jiang et al. [5]	Our method
1	0.64	**0.81**
2	**0.72**	0.71
4	0.53	**0.65**
6	0.72	**0.82**
7	**0.68**	0.64
10	**0.66**	0.60
12	0.74	**0.79**
15	0.53	**0.67**
17	0.70	**0.73**
18	**0.75**	0.72
25	0.57	**0.60**
26	0.53	**0.59**
Mean	0.65	**0.69**

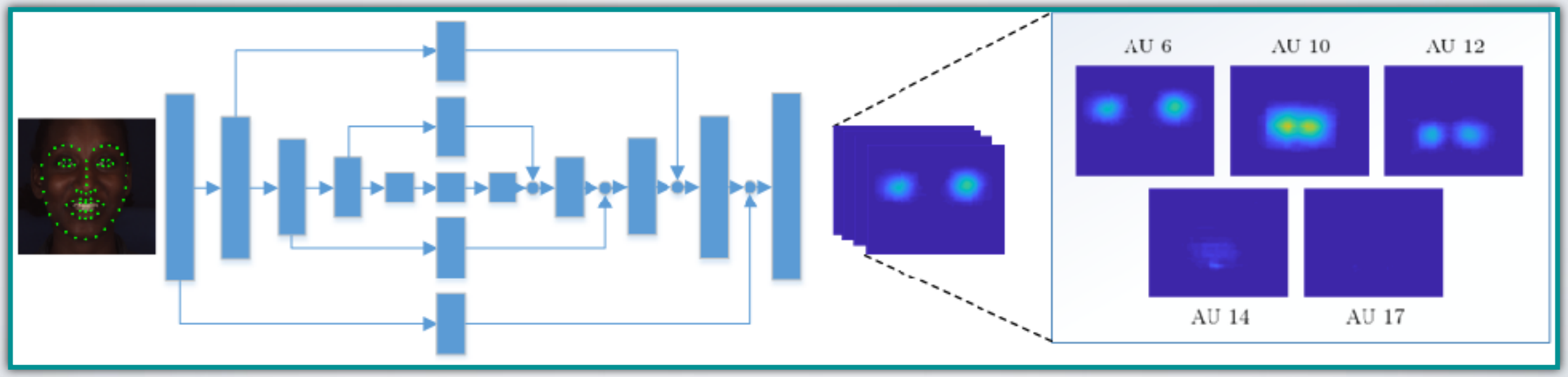
Jaiswal, Martinez, Valstar (2015), *'Learning to combine local models for Facial Action Unit detection'*, Proc. IEEE Int'l Conf. Face and Gesture Recognition (FG)

Break

FACS Action Unit: Smile

AU HEATMAP REGRESSION





AU HEATMAP REGRESSION

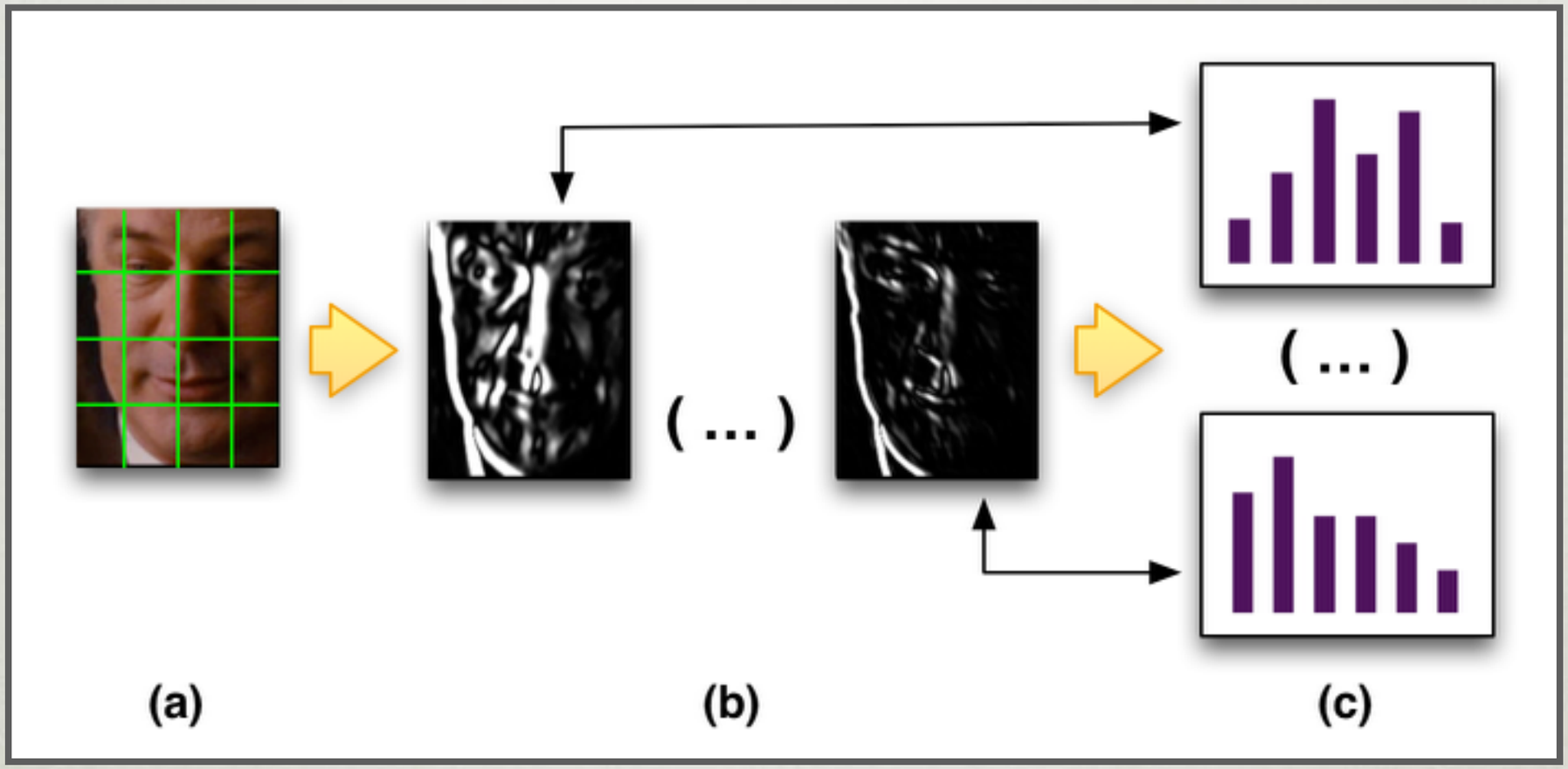
	AU	6	10	12	14	17	Avg.
ICC	Our	**0.79**	**0.80**	0.86	**0.54**	0.43	**0.68**
	Single Heatmap	0.78	0.79	0.84	0.36	0.49	0.65
	ResNet18	0.71	0.76	0.84	0.43	0.44	0.64
	2DC	0.76	0.71	0.85	0.45	**0.53**	0.66
	CCNN-IT	0.75	0.69	0.86	0.40	0.45	0.63
	VGP-AE	0.75	0.66	**0.88**	0.47	0.49	0.65
MSE	Our	0.77	0.92	**0.65**	1.57	**0.77**	**0.94**
	Single Heatmap	0.89	1.04	0.82	2.24	0.78	1.15
	ResNet18	0.98	**0.90**	0.69	1.88	0.95	1.08
	2DC	**0.75**	1.02	0.66	**1.44**	0.88	0.95
	CCNN-IT	1.23	1.69	0.98	2.72	1.17	1.57
	VGP-AE	0.82	1.28	0.70	**1.43**	**0.77**	1.00

E. Sanchez-Lozano, G. Tzimiropoulos and M.F. Valstar (2018). *"Joint Action Unit localisation and intensity estimation through heatmap regression"*, BMVC 2018

Temporal Features

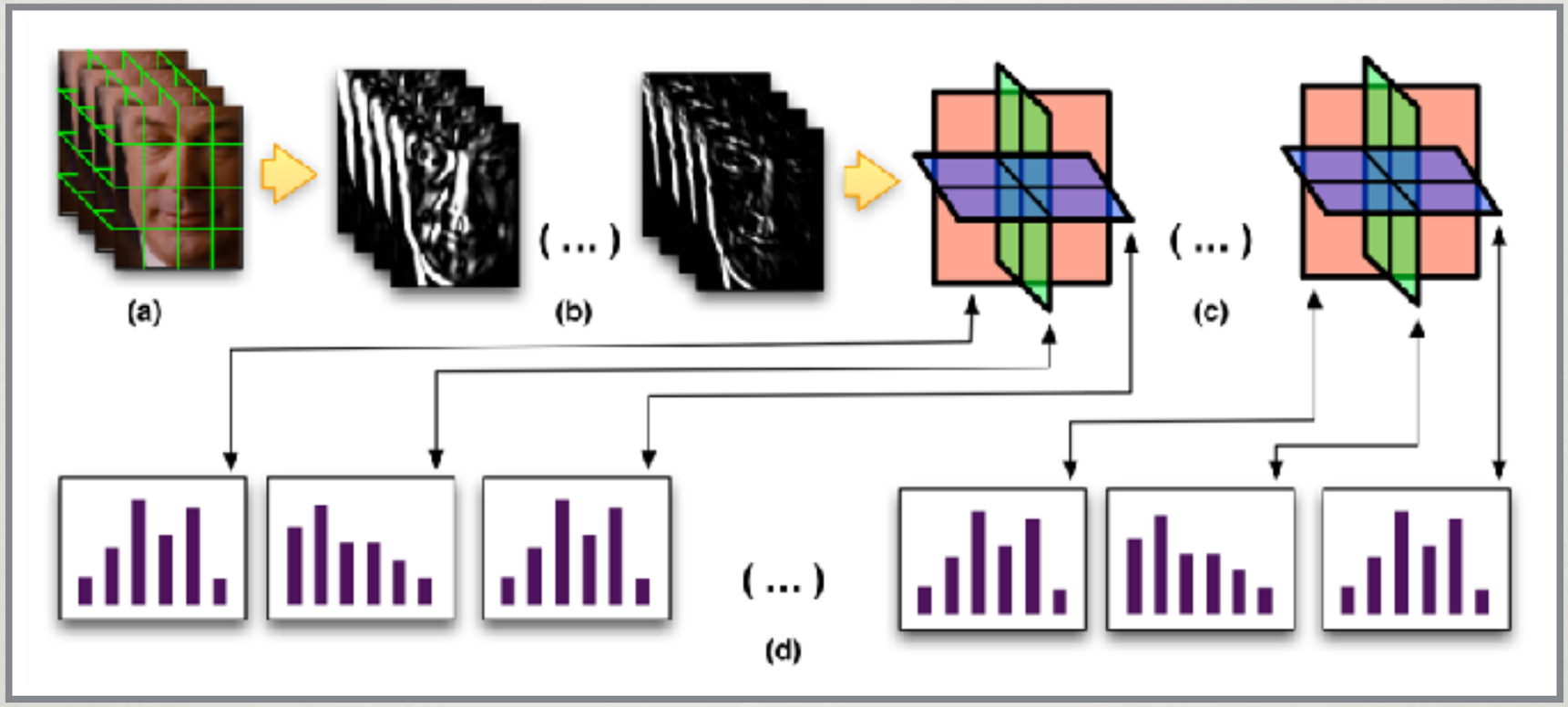
- Appearance Features describe static face appearance
- But expression recognition is largely action recognition!

Local Gabor Binary Patterns



- Gabor Filters applied to create Gabor Pictures
- Local Binary Patterns applied on Gabor Pictures
- Attains higher accuracy than either applied separately

LGBP-TOP



- To model changes in appearance Volume Local Binary Patterns were introduced
- Proved problematic due to large feature dimensionality
- LBP-Three Orthogonal Planes calculates LBP in xy, xt, and yt planes only
- We extended LGBP to LGBP-TOP in a similar fashion

LGBP-TOP results

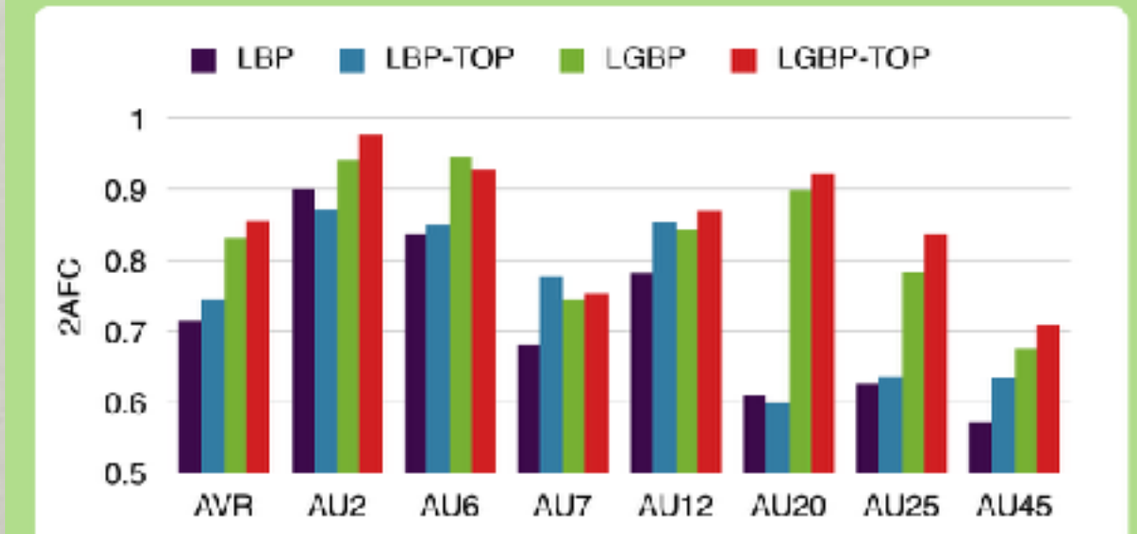


Figure 2: MMI database performance

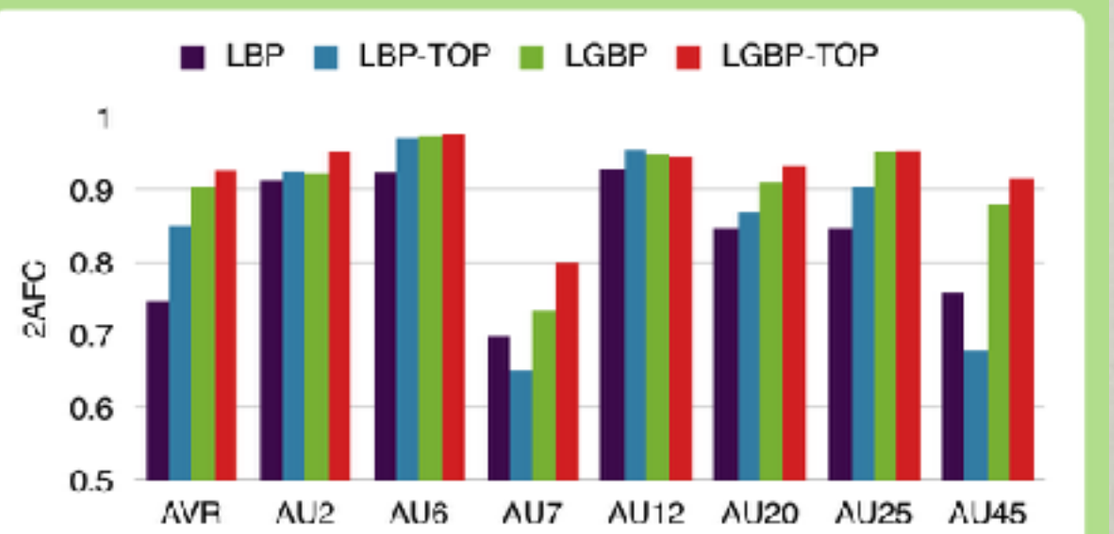
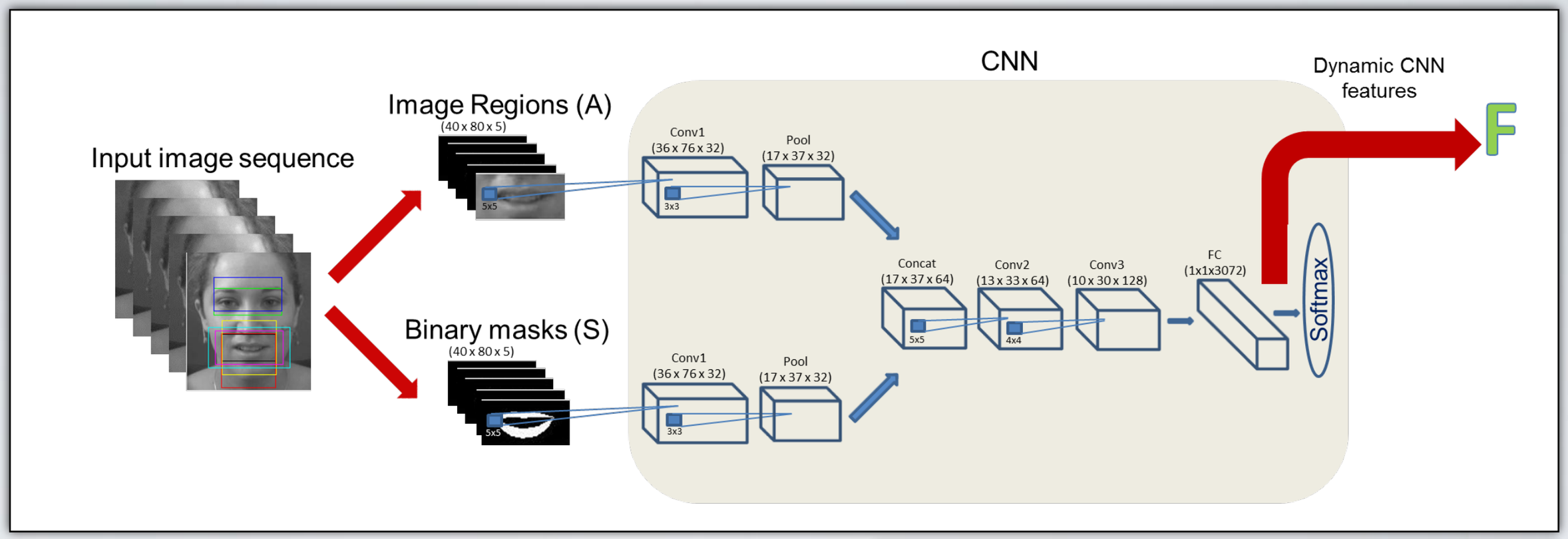


Figure 3: Cohn-Kanade database performance

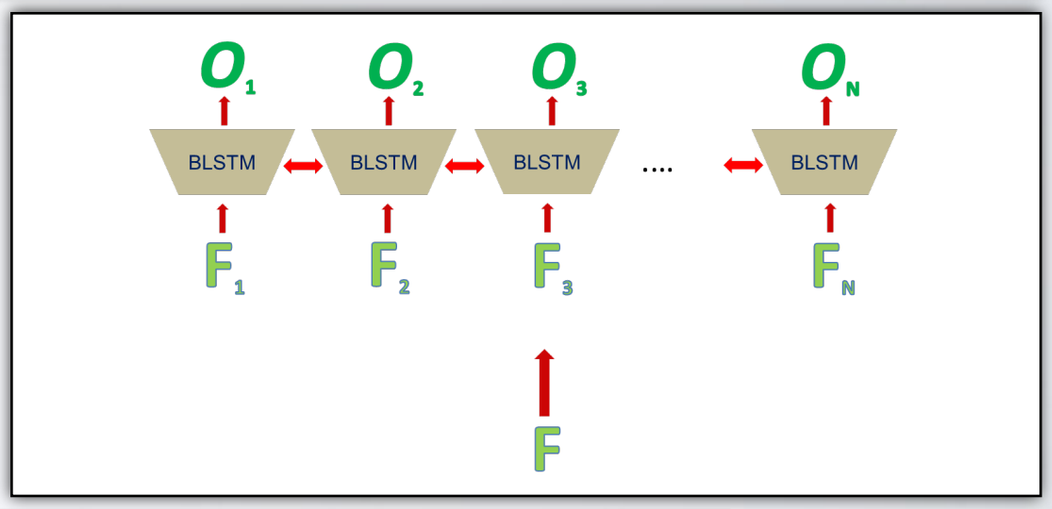
Almaev & Valstar (2013), *'Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition'*, Proc. Affective Computing and Intelligent Interaction (ACII'13)

DYNAMIC DEEP LEARNING



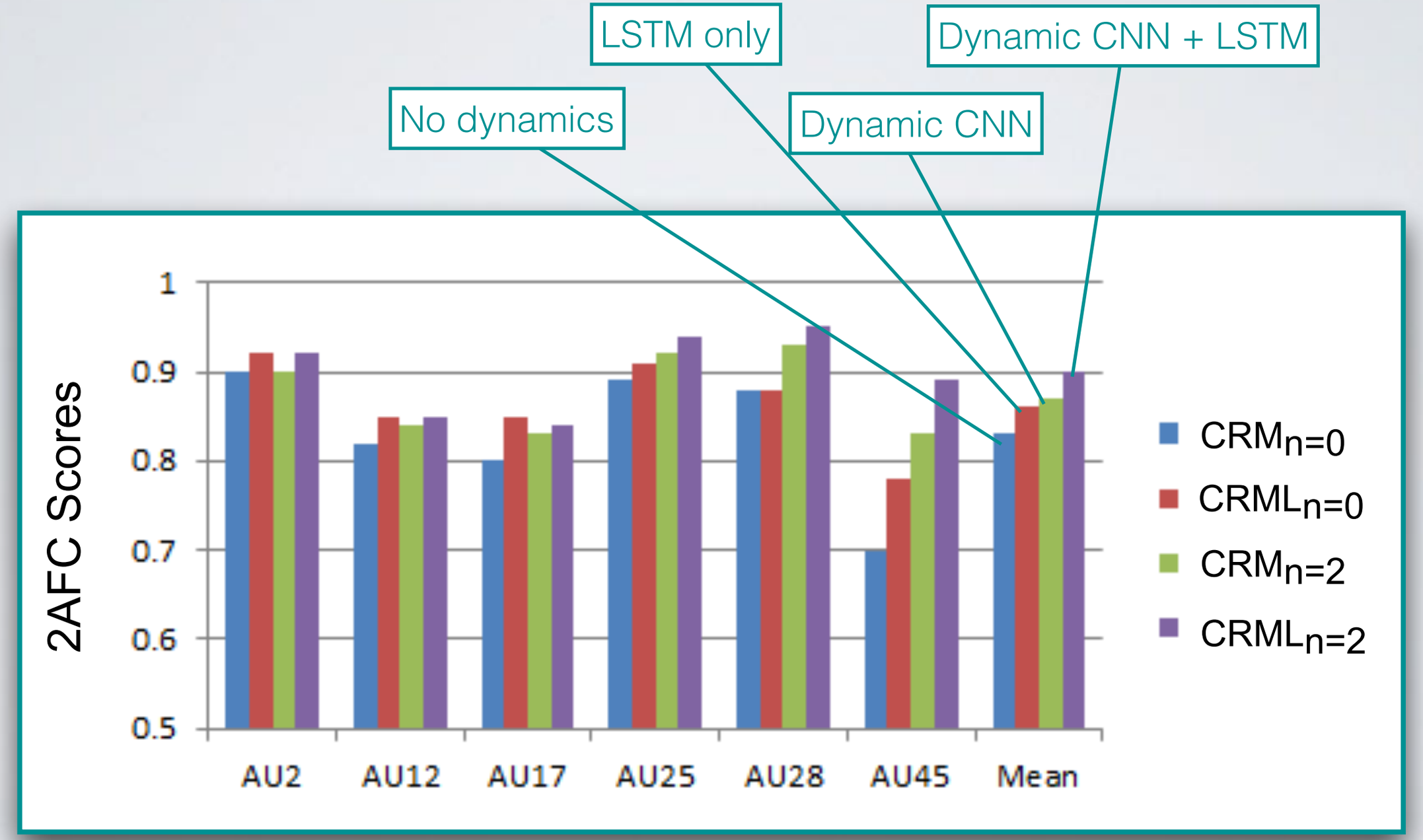
Dynamics are encoded by:

- Dynamic CNN
- Image difference
- Long-Short Term Memory Neural Networks



S. Jaiswal and M.F. Valstar (2016). *"Deep Learning the Dynamic Appearance and Shape of Facial Action Units"*, WACV [Code available!]

EFFECT OF DYNAMICS



B. Martinez, M.F. Valstar, B. Jiang and M. Pantic (2017). *"Automatic Analysis of Facial Actions: a Survey"*, IEEE Trans. Affective Computing

FERA 2015 PERFORMANCE

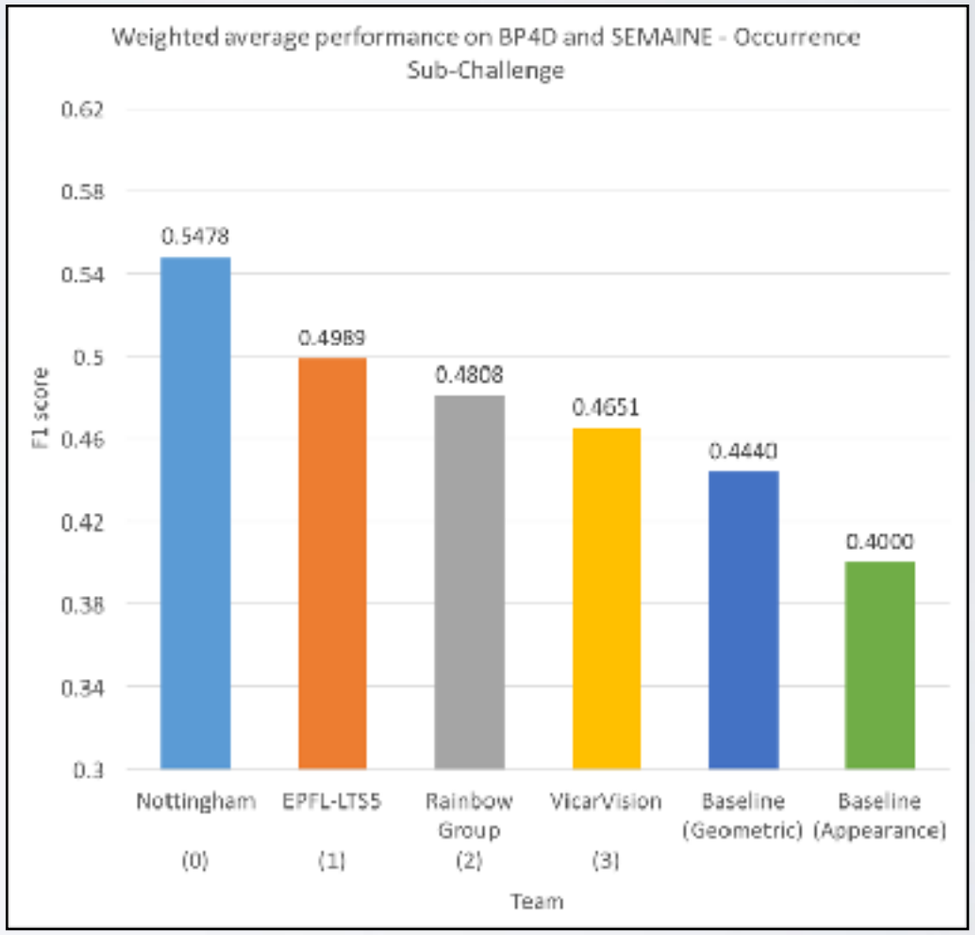
AU	LGBP [22]	GDNN	DLE [27]	CRML$_{n=2}$
2	0.75	0.67	0.66	0.80
12	0.52	0.63	0.76	0.74
17	0.07	0.14	0.25	0.32
25	0.40	0.77	0.61	0.85
28	0.01	0.31	0.26	0.33
45	0.21	0.55	0.35	0.57
Mean	0.33	0.51	0.48	0.60

Table 1. Performance (F1 scores) comparison on SEMAINE test set.

AU	LGBP [22]	GDNN	DLE [27]	CRML$_{n=2}$
1	0.18	0.33	0.25	0.28
2	0.16	0.25	0.17	0.28
4	0.22	0.21	0.28	0.34
6	0.67	0.64	0.73	0.70
7	0.75	0.79	0.78	0.78
10	0.80	0.80	0.80	0.81
12	0.79	0.78	0.78	0.78
14	0.67	0.68	0.62	0.75
15	0.14	0.19	0.35	0.20
17	0.24	0.28	0.38	0.36
23	0.24	0.33	0.44	0.41
Mean	0.44	0.48	0.51	0.52

Table 2. Performance (F1 scores) comparison on BP4D Test set.

Occurrence detection



S. Jaiswal and M.F. Valstar (2016). *"Deep Learning the Dynamic Appearance and Shape of Facial Action Units"*, WACV [Code available soon!]

Machine Learning for Expression Recognition

- Very large number of methods proposed
  - SVM
  - HMM
  - ANN
  - CRF
  - Gaussian Processes
  - RVM
- Architectures depend on the problem

Class Types

	Multiclass	Multiple Independent Binary	Regression
Discrete Emotions	✓		
Action Units (Occurrence)		✓	
Action Units (Intensity)		✓	✓
Action Units (Temporal Phases)		✓	
Dimensional Affect			✓
Higher-order	✓	✓	✓

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

BEHAVIOMEDICS

AC/SSP

- Affective Computing:

> "…[is] computing that relates to, arises from, or deliberately influences emotion or other affective phenomena." (Rosalind Picard)

- Social Signal Processing:

> "…[is a] research and technological domain that aims at providing computers with the ability to sense and understand human social signals." (Vinciarelli et al.)

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

DEFINITION

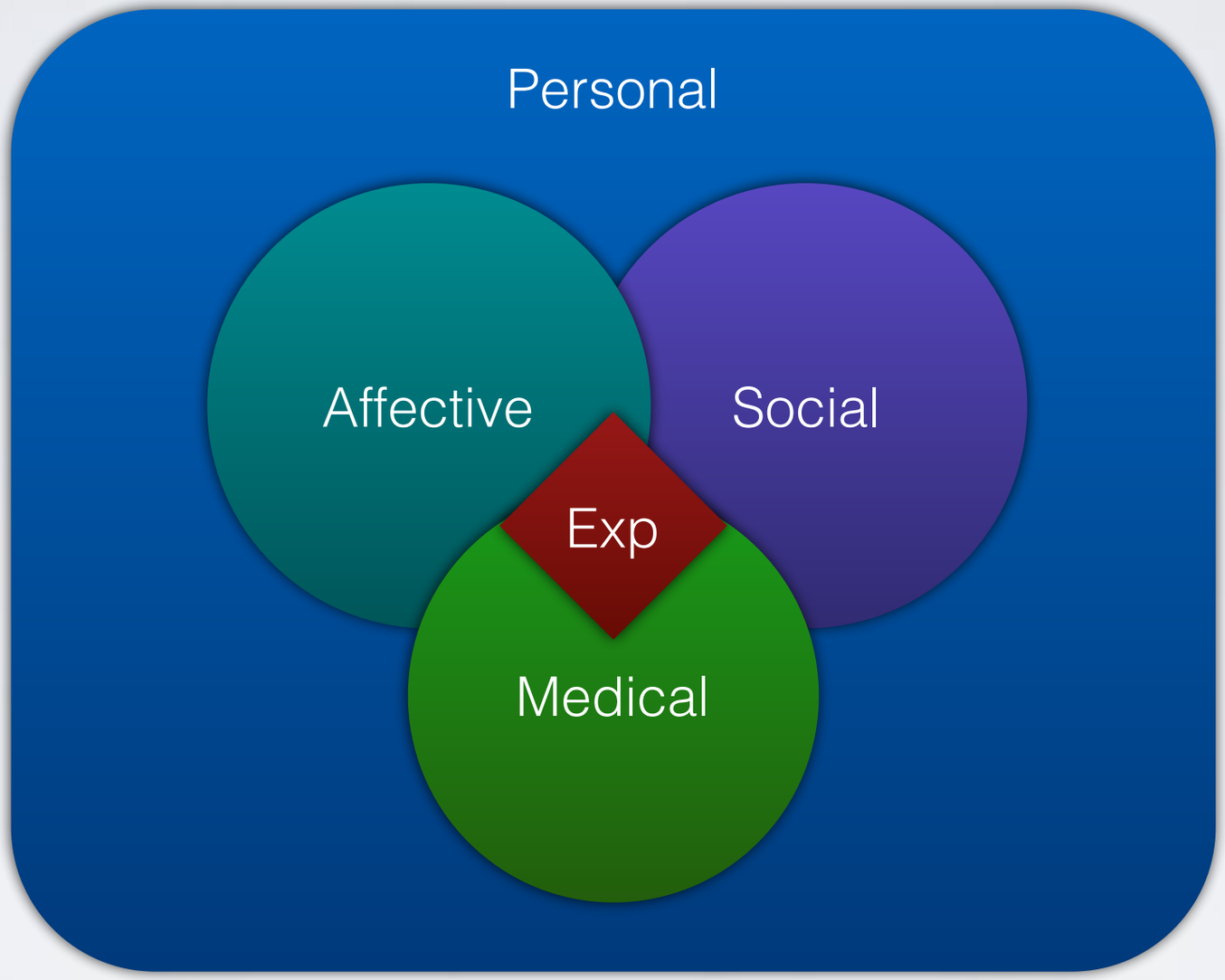
Behaviomedics - The application of automatic analysis and synthesis of affective and social signals to aid objective diagnosis, monitoring, and treatment of medical conditions that alter one's affective and socially expressive behaviour.

M.F. Valstar, "*Automatic Behaviour Understanding in Medicine",* proc. Int'l Conference Multimodal Interaction (workshop section), 2014

PROPOSITION

- Great opportunities for the application of existing AC/SSP methods in medicine
- Novel diagnosis, monitoring, and treatment options will have massive societal impact
- Feed-back to AC/SSP with new data and challenges to overcome

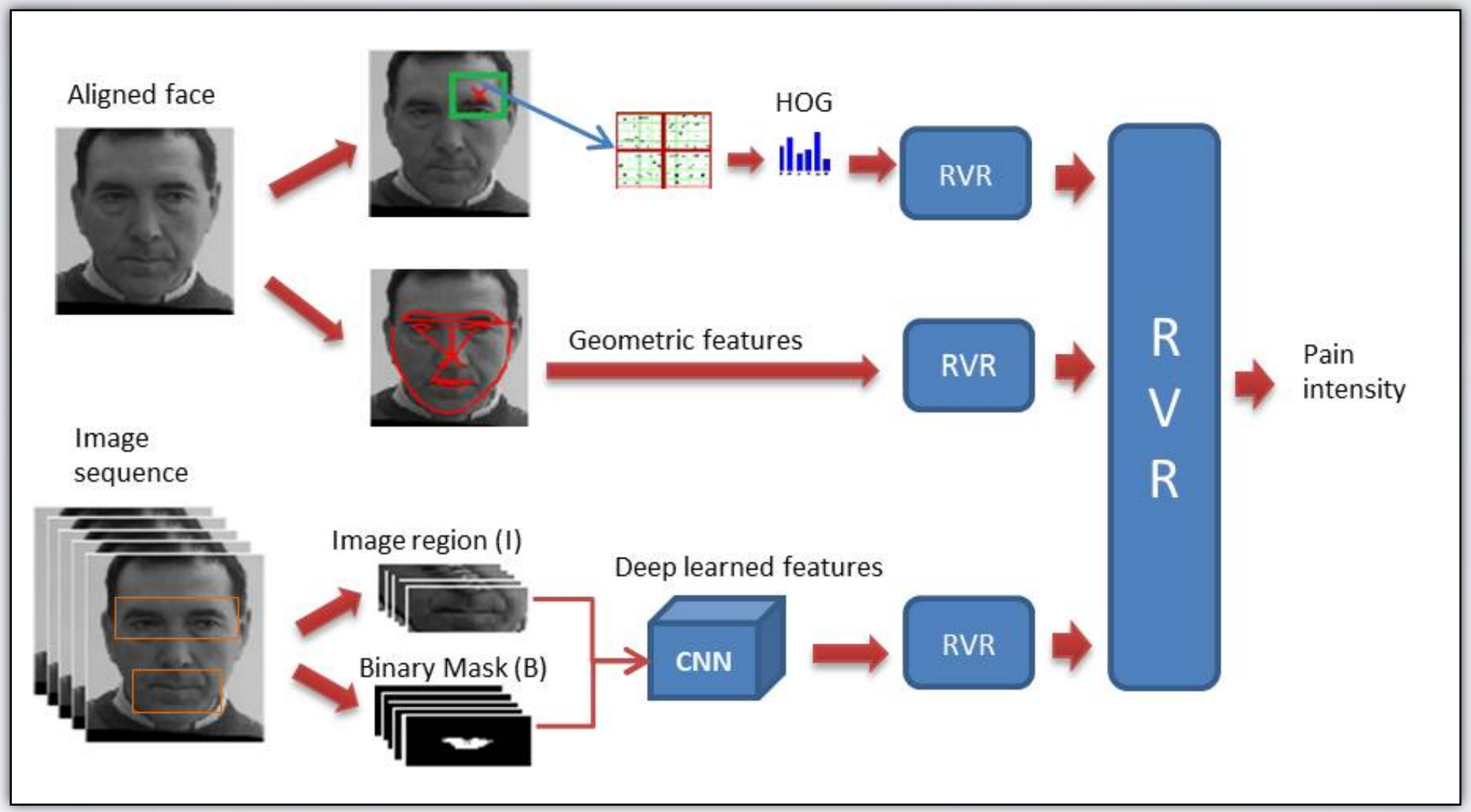
INFERENCE OF MEANING



PROBLEMS WITH CNNS

- Need large amount of data
- Still relatively slow
- When you work with small datasets, hand-crafted features may well outperform (deep) learned features

PAIN ESTIMATION



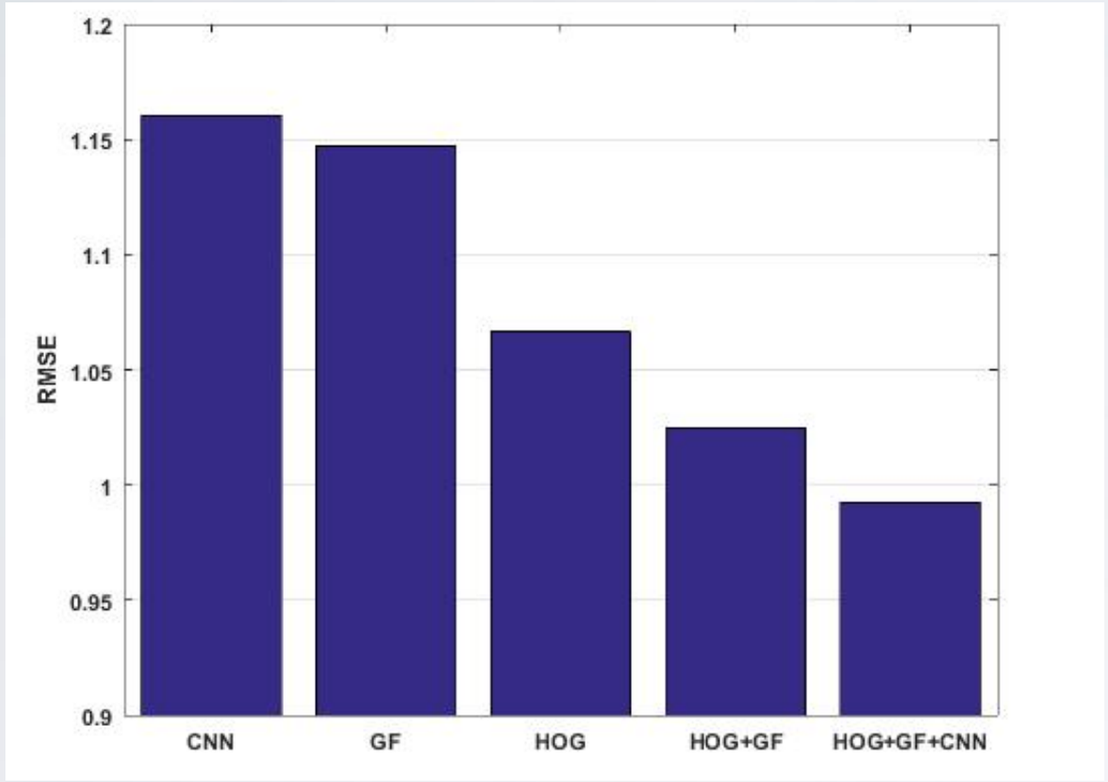
PAIN ESTIMATION

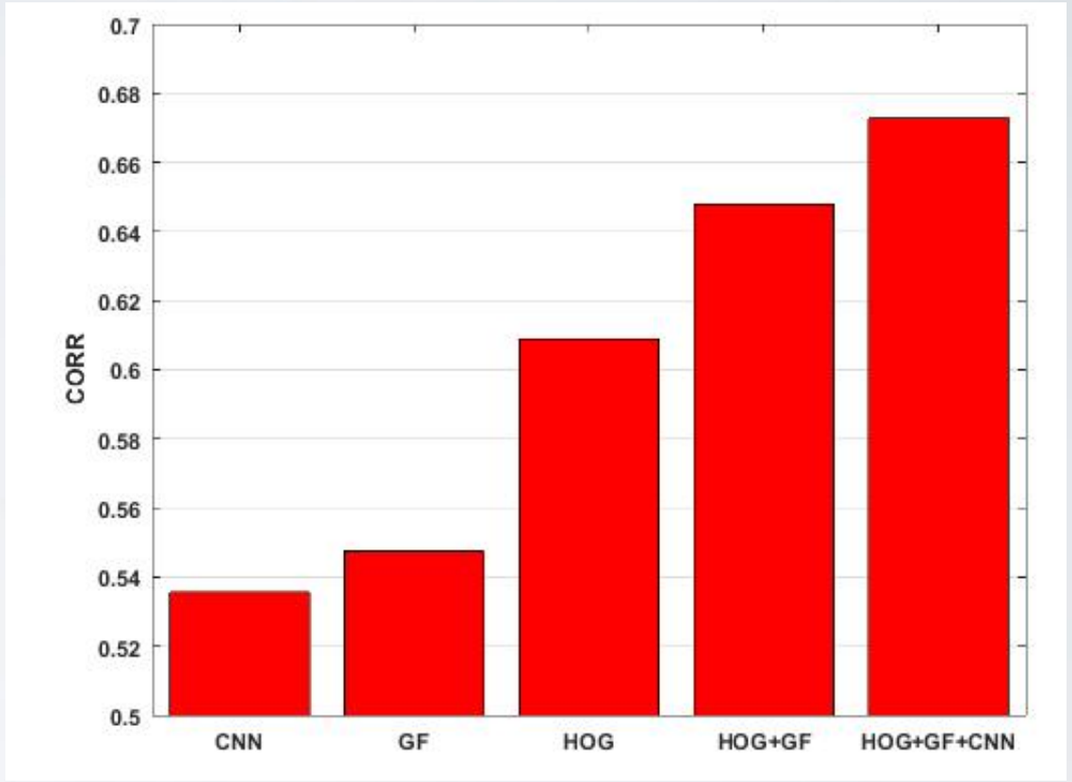
	RMSE	CORR
Kaltwang et.al	1.1773	0.59
Neshov and Manolava	1.1314	0.59
Kaltwang et.al	1.6911	0.66
Our method	**0.9926**	**0.6728**

State of the art results obtained on the UBC McMaster database for both RMSE and Correlation

J. Egede, M.F. Valstar, and B. Martinez, "*Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation ",* Face and Gesture Recognition, 2017

PAIN ESTIMATION

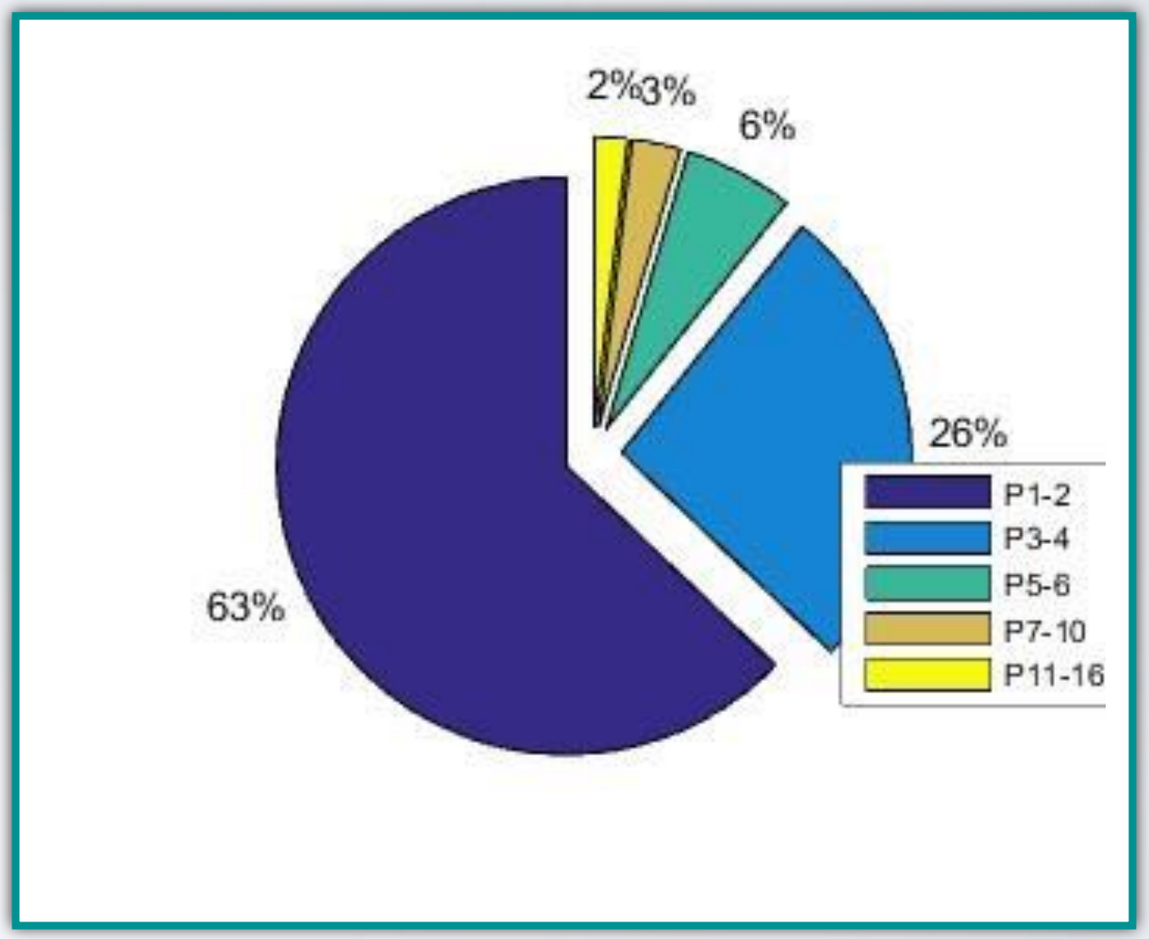




- Fusing learned and handcrafted features improves both RMSE and Correlation
- Geometric and appearance features both help (CNN has both)

HIGHLY UNBALANCED DATA

- 80 % of frames are neutral
- Non-neutral frames are also highly unbalanced
- Only 0.85% of total frames has pain levels > 6!

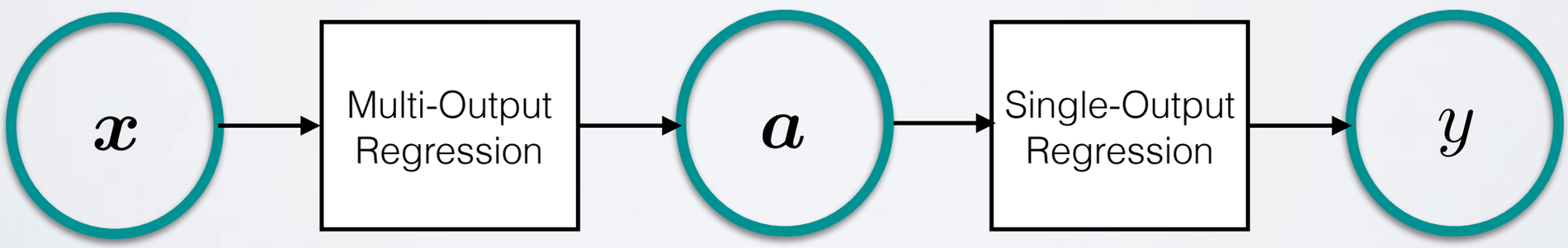


CUMULATIVE ATTRIBUTES

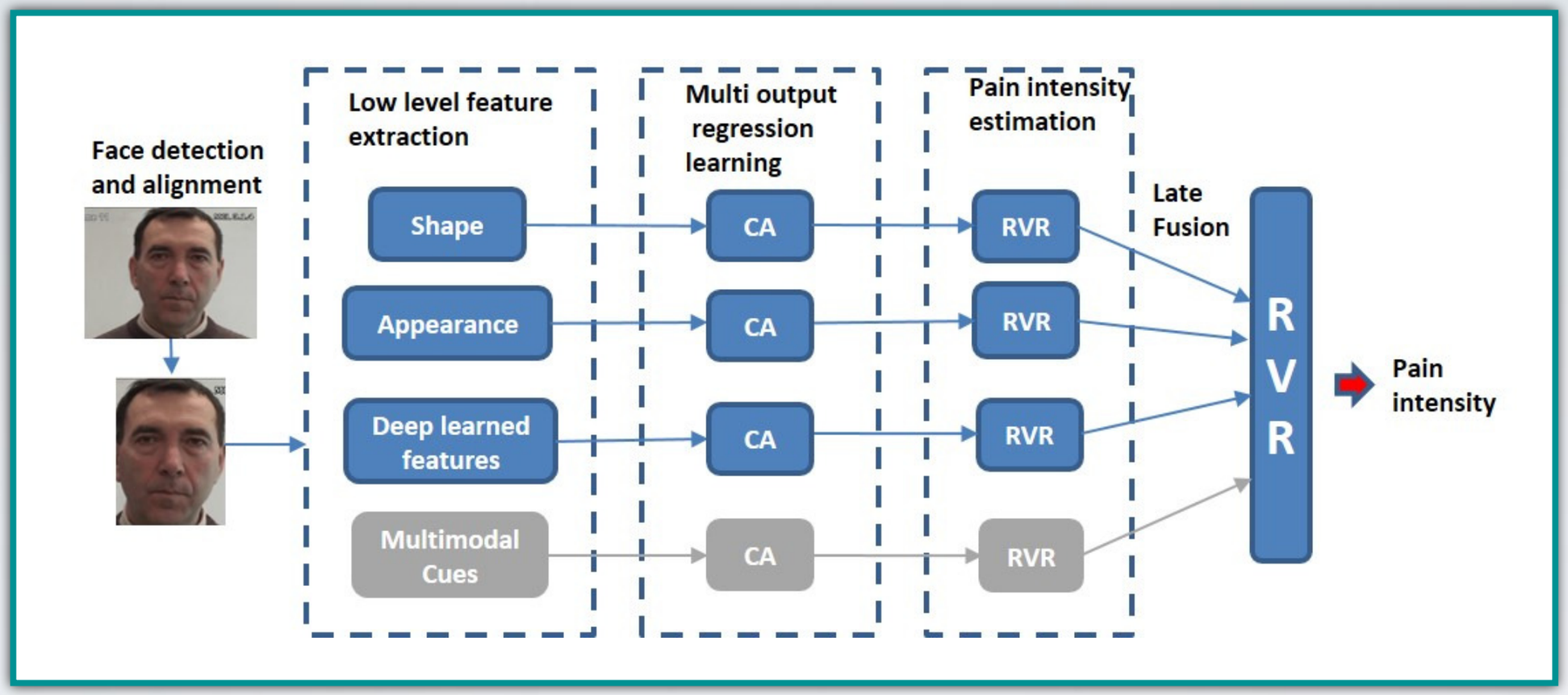
- For an ordinal classification problem of k levels, we create an intermediate attribute space $\mathbf{a}$ of dimension k-1 as follows:

$$a_t^j = \begin{cases} 1, & \text{when } j \leq y_t \\ 0, & \text{when } j > y_t \end{cases}$$

E.g. if y = 3 and k = 10:
$\mathbf{a} = [1, 1, 1, 0, 0, 0, 0, 0, 0]$

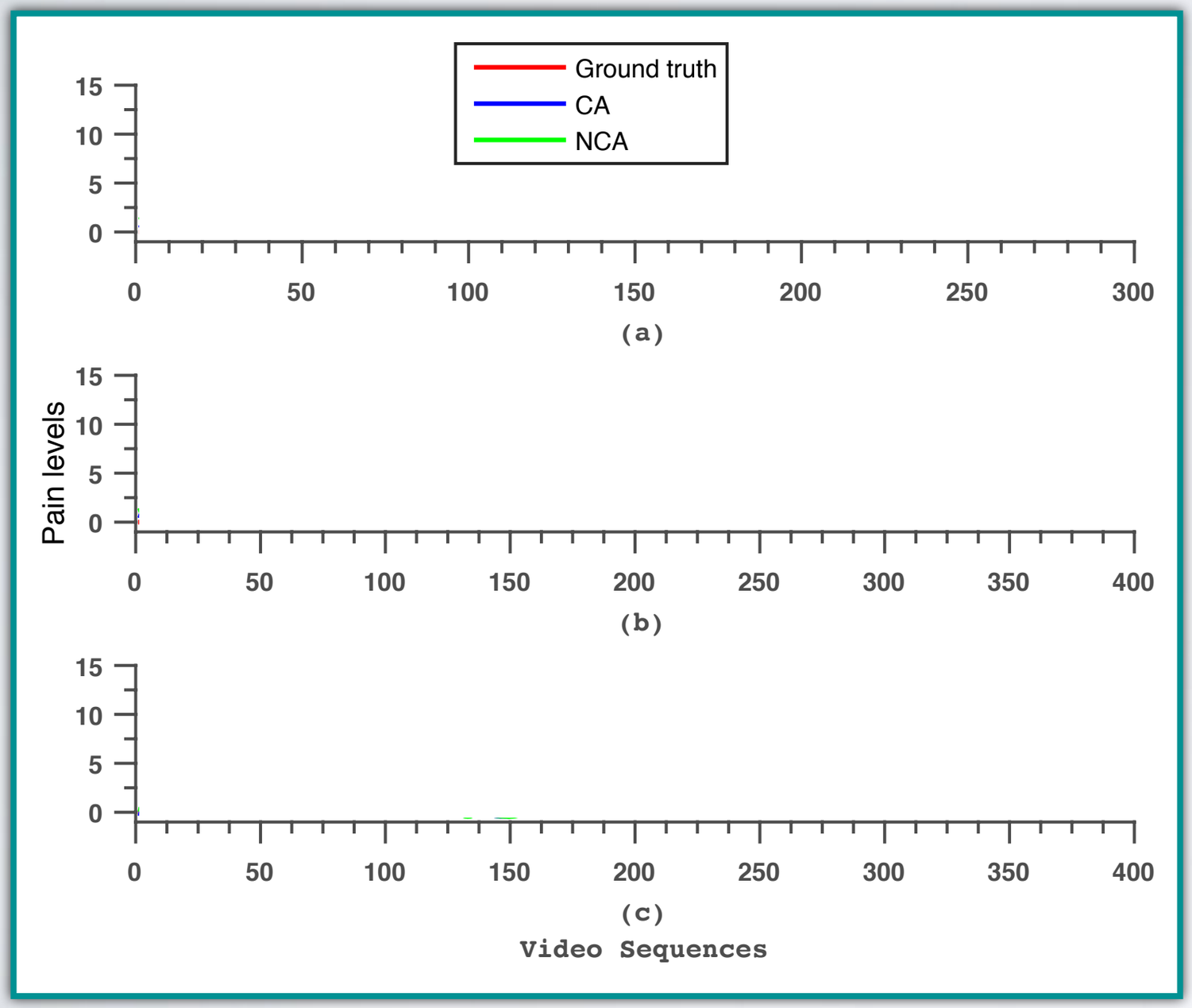


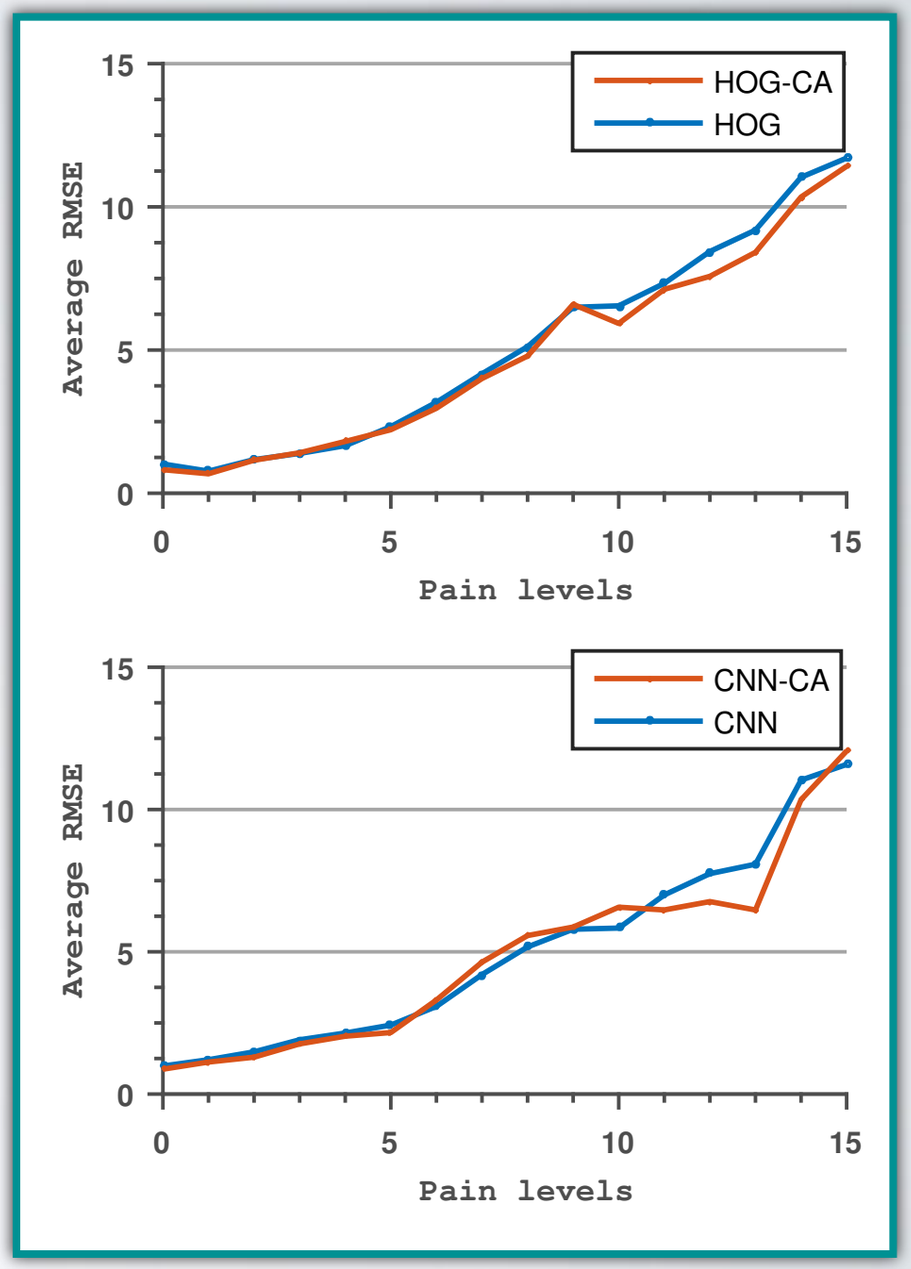
CUMULATIVE ATTRIBUTES



- Cumulative attributes first proposed for age estimation
- Ideal for ordinal problems where higher value classes share attributes of lower value classes

CA WORKS


Ground truth
CA
NCA
(a)
Pain levels
(b)
(c)
Video Sequences

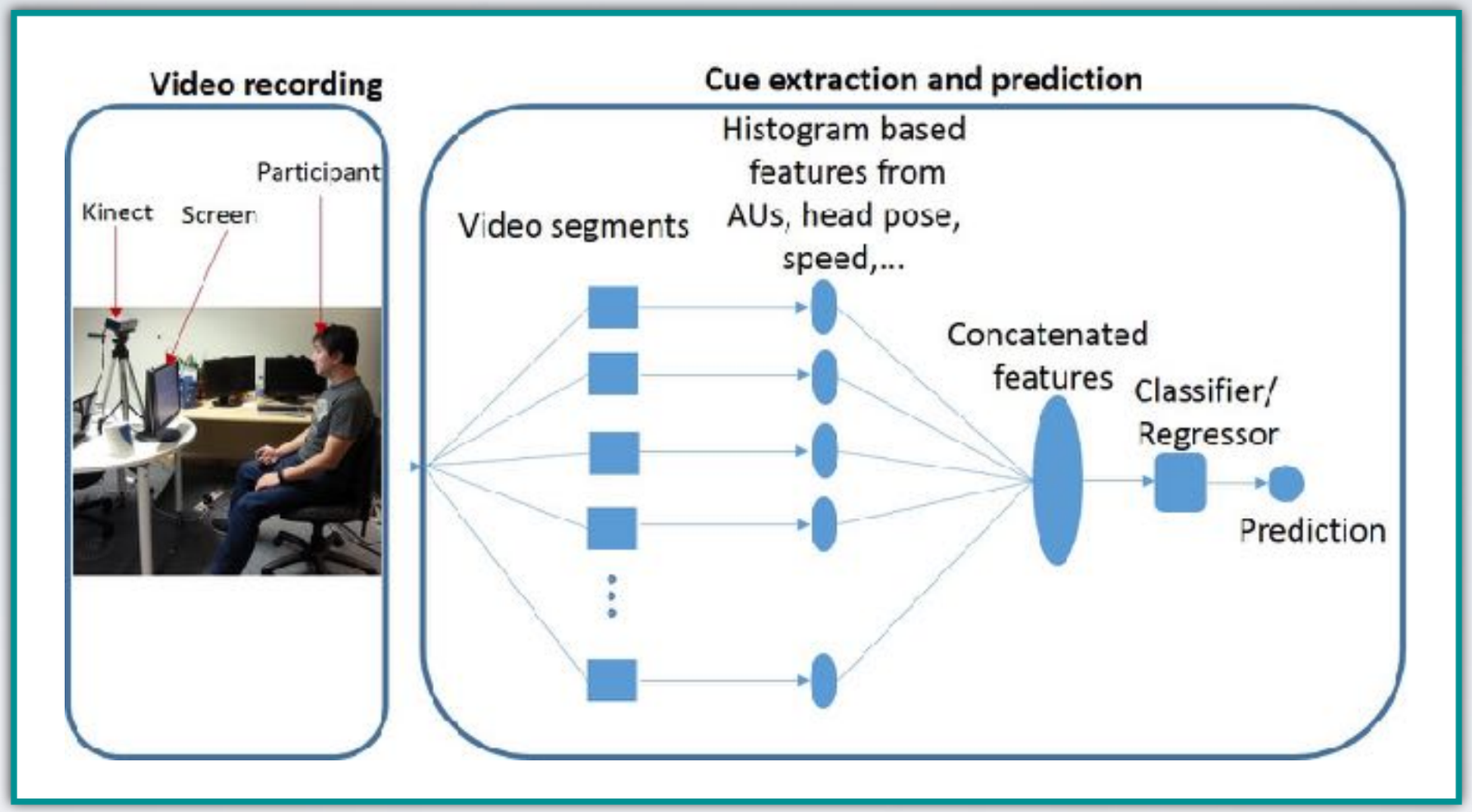

HOG-CA
HOG
Average RMSE
Pain levels
CNN-CA
CNN
Average RMSE
Pain levels

CA RESULTS

	RMSE	PCC
Kaltwang et al. [17]	1.18	0.59
Neshov & Manolava [29]	1.13	0.59
Kaltwang et al. [18]	1.69	0.66
Florea et al. [12]	1.10	0.53
Zhou et al. [46]	1.24	0.65
Egede et al. [8]	1.30	0.63
Our method	**1.04**	**0.64**

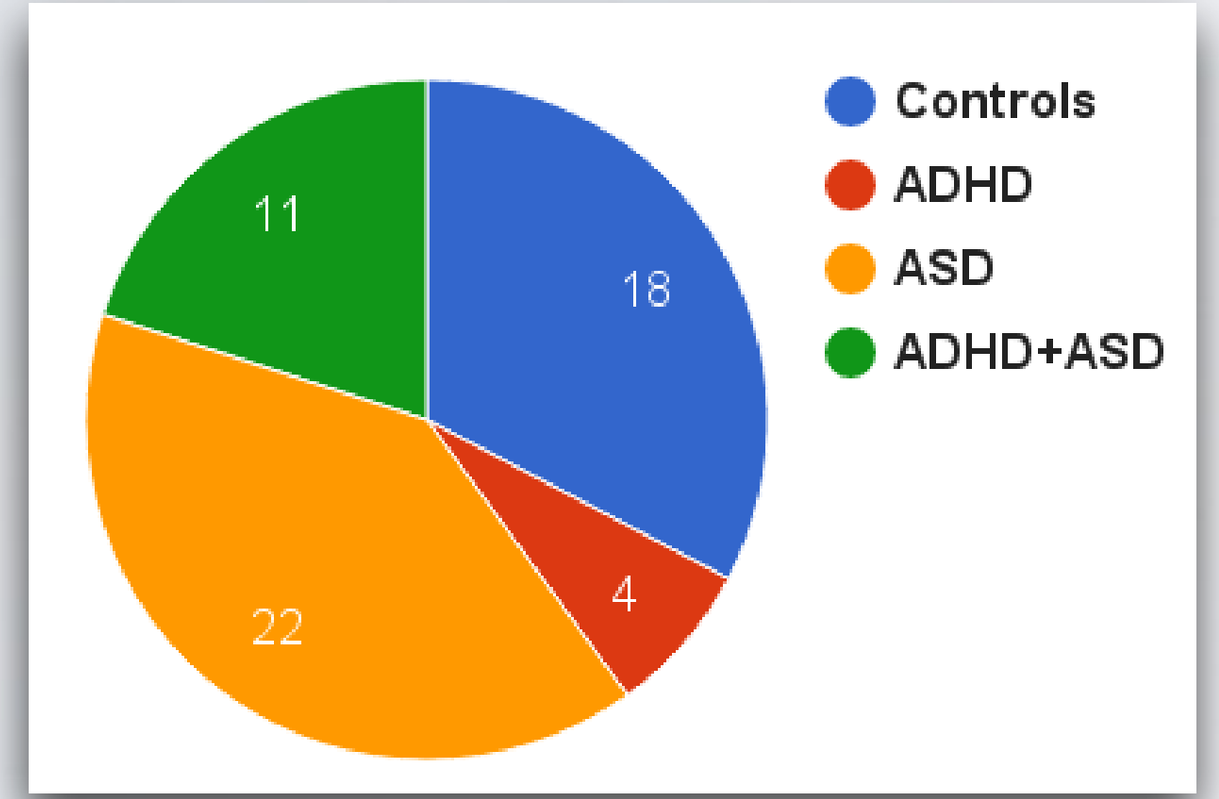
J. Egede and M.F. Valstar, "*Cumulative Attributes for Pain Intensity Estimation",* proc. Int'l Conference Multimodal Interaction, 2017

ADHD/ASD



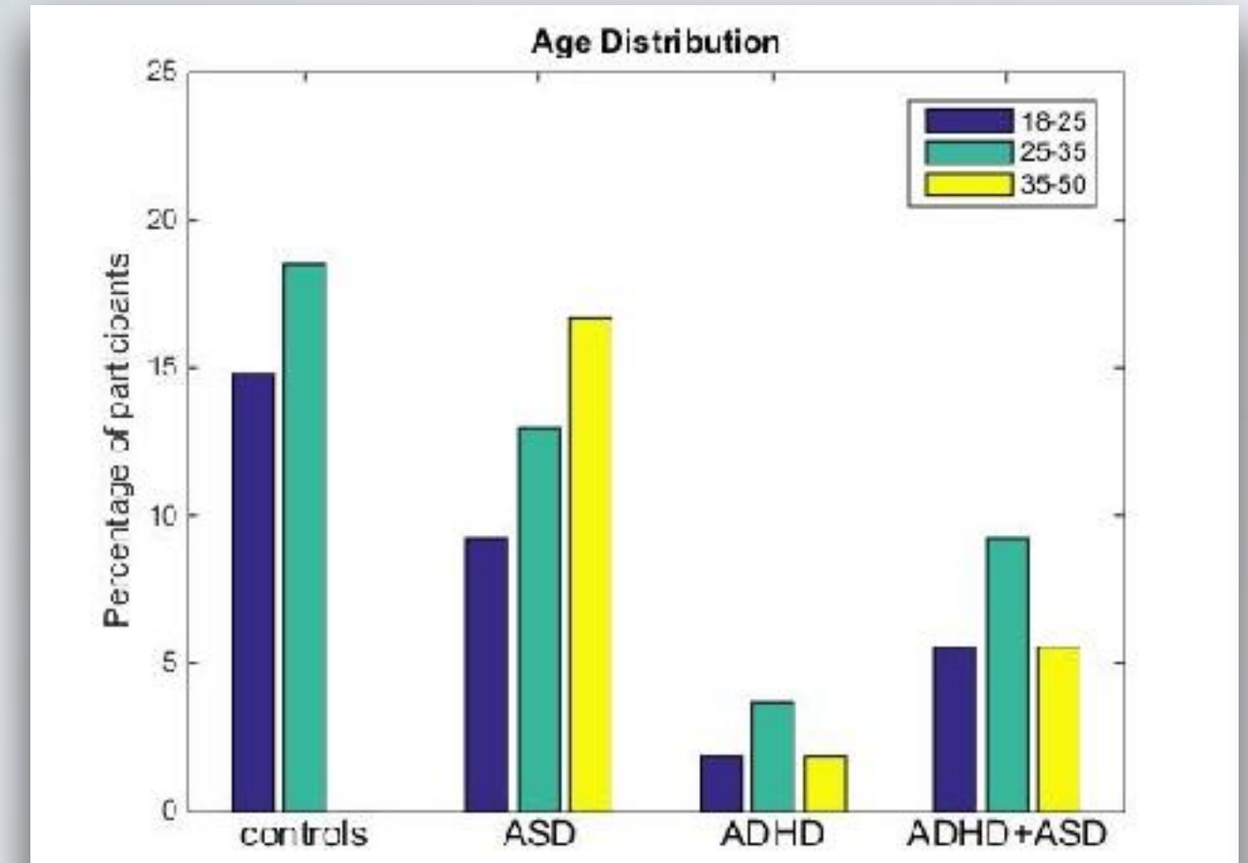
- Patients from a clinic and controls were recorded by a Kinect while answering a series of questions of the standard 'Strange Stories' questionnaire
- Dynamic Deep Learned AUs, head pose, and derived features per question were collected for prediction of ADHD/ASD: 2652 features (221/question)

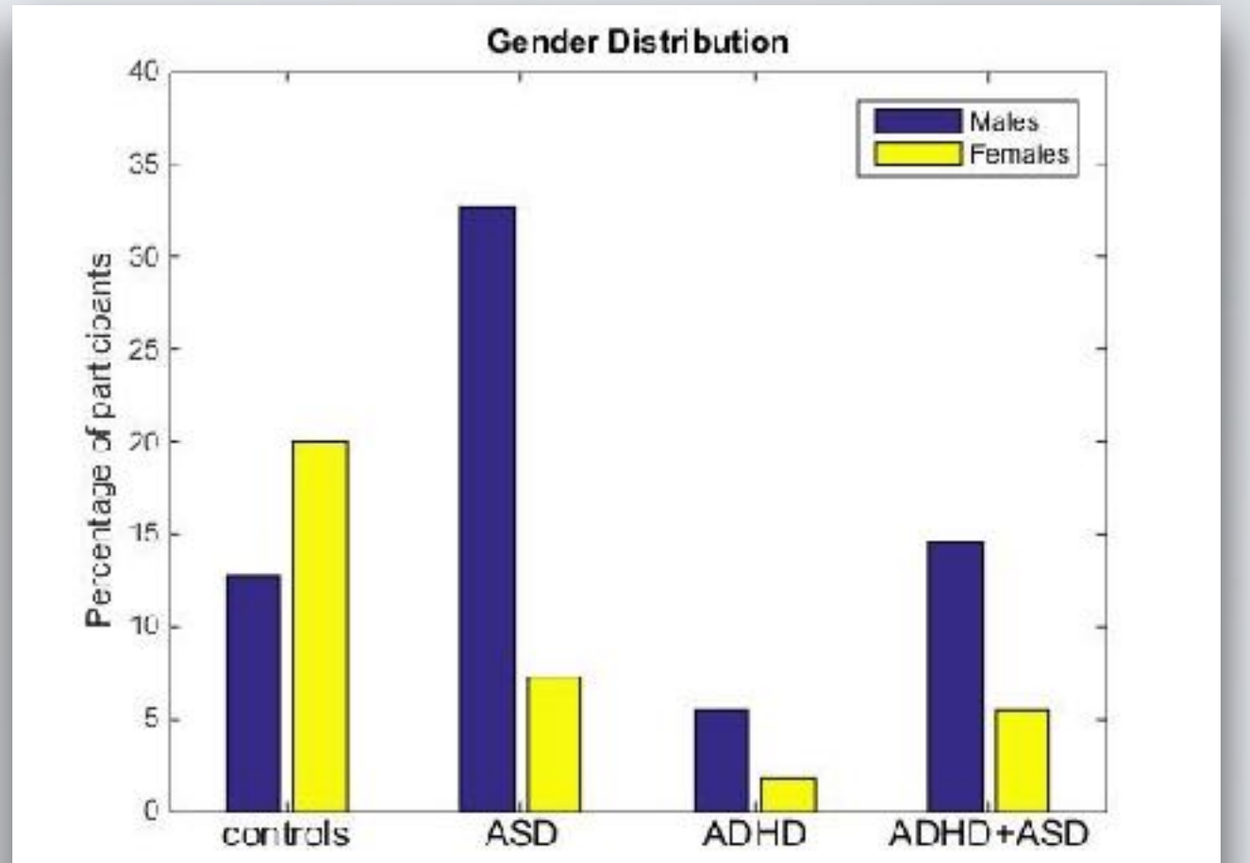
ADHD/ASD DATA



- Participants with ADHD/ASD were recruited from the local 'Asperger Clinic'
- Their condition labels were those assigned by clinicians
- Controls were recruited from the local population

ADHD/ASD DATA





- Control group needs to be adjusted to include more men and older people
- We need more participants with ADHD only

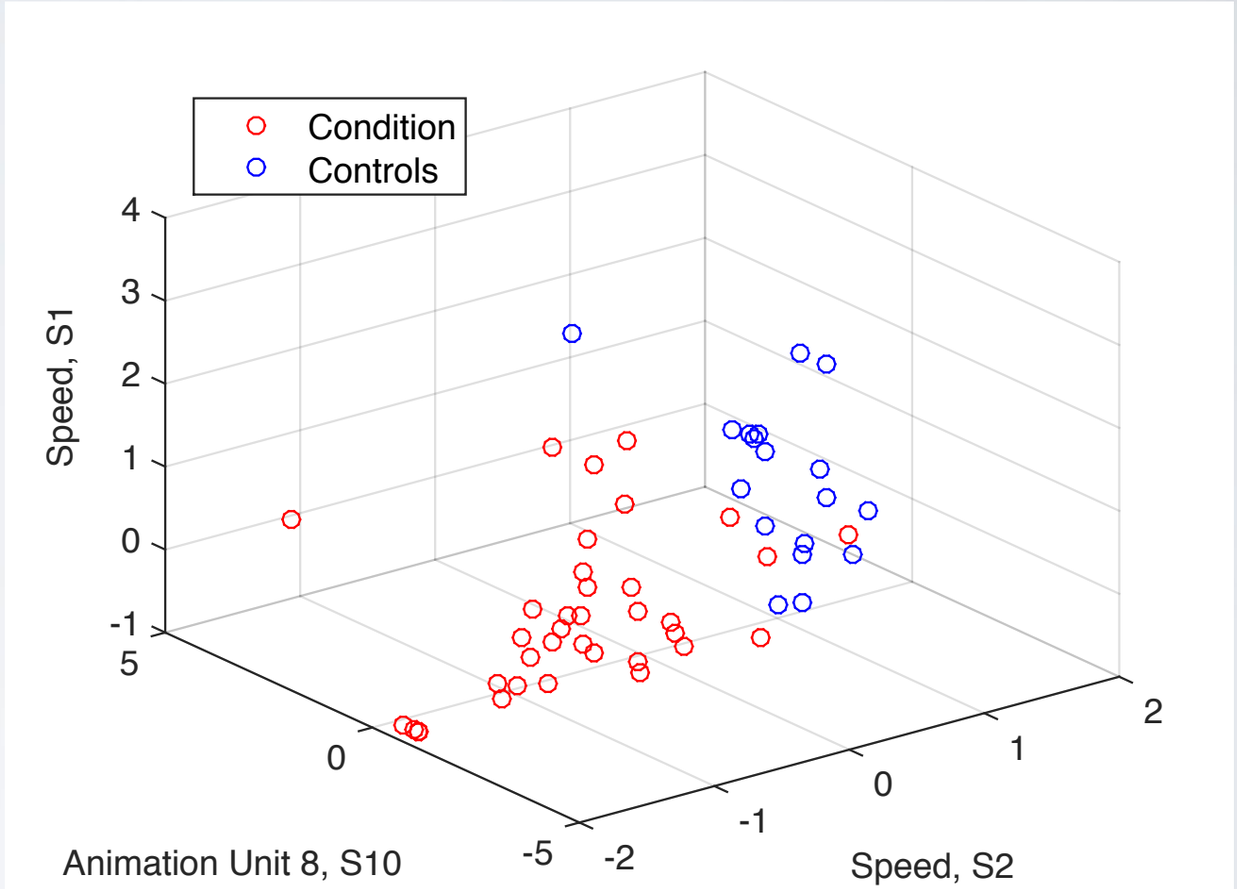
BEHAVIOUR PRIMITIVES

- Seven AUs were detected: AU1, AU2, AU4, AU12, AU15, AU20, AU45 together with 12 Kinect Animation Units (AnU
- Histograms of AU/AnU intensities were computed over all the frames in a video segment
- One 10-bin histogram was computed for each AU
- For AU45, the frequency of its occurrence and the average duration of its activation were estimated in each video segment.
- Head rotation, head speed, and head displacement were calculated in 3D

CONTROLS VS CONDITION

	Correct	Incorrect
Controls	16	2
Condition	37	0

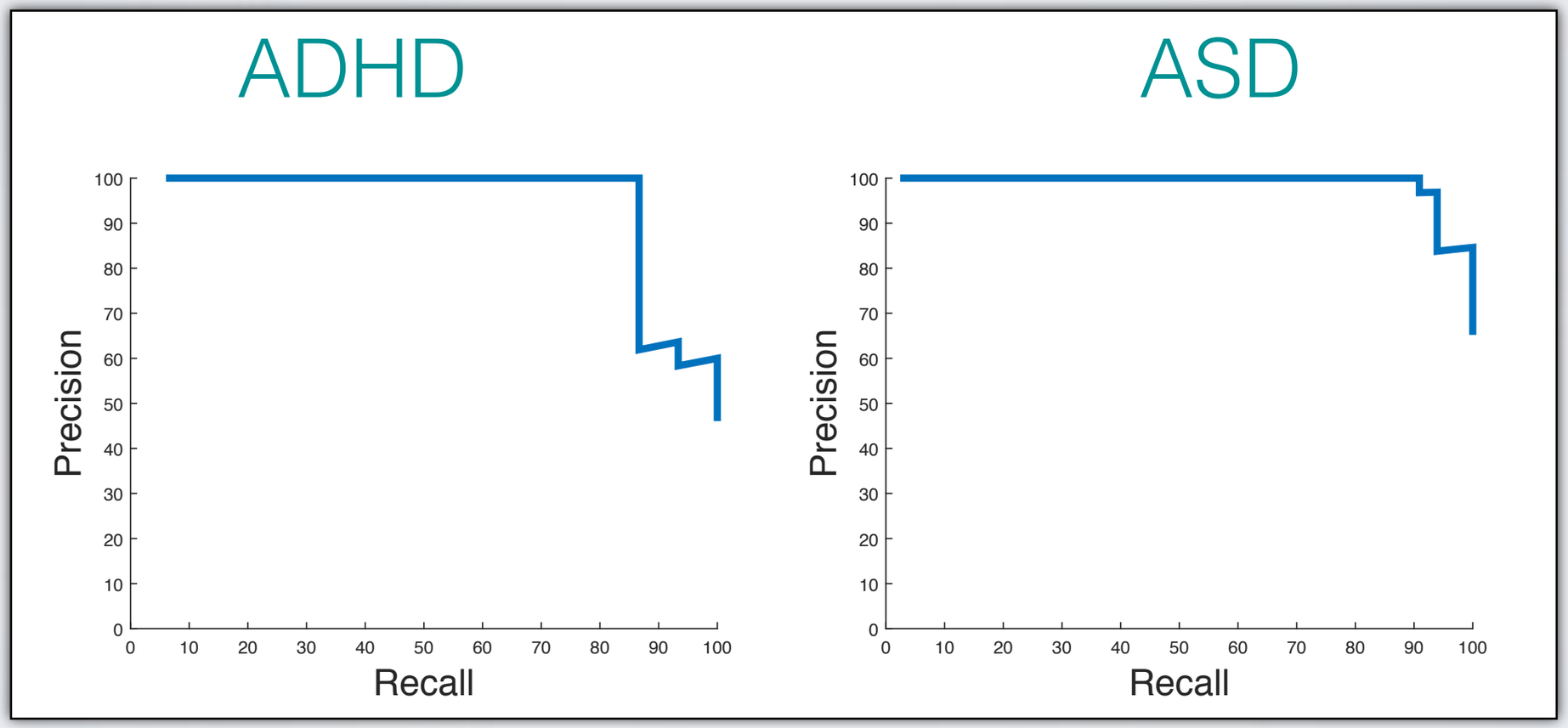
- Forward feature selection applied
- Classification using Support Vector Machines with a Radial Basis Function kernel



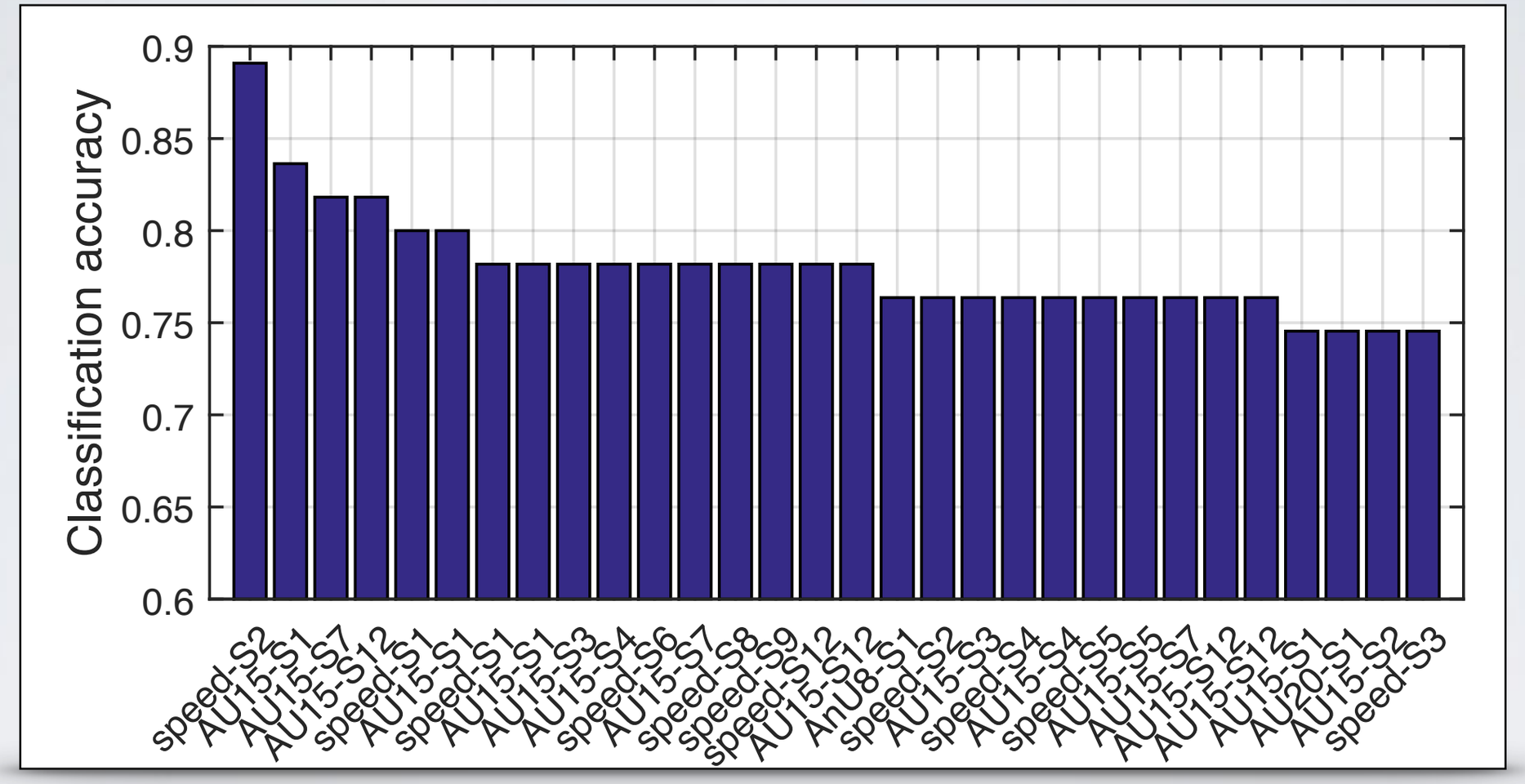
- Classification rate is 96.4% correct! Only two errors made in Leave-One-Out cross-validation

ADHD/ASD VS CONTROLS

Classifier	Accuracy	Precision	Recall	AUC
ADHD vs Controls	0.91	1.0	0.80	0.95
ASD vs Controls	0.88	0.8	0.94	0.99



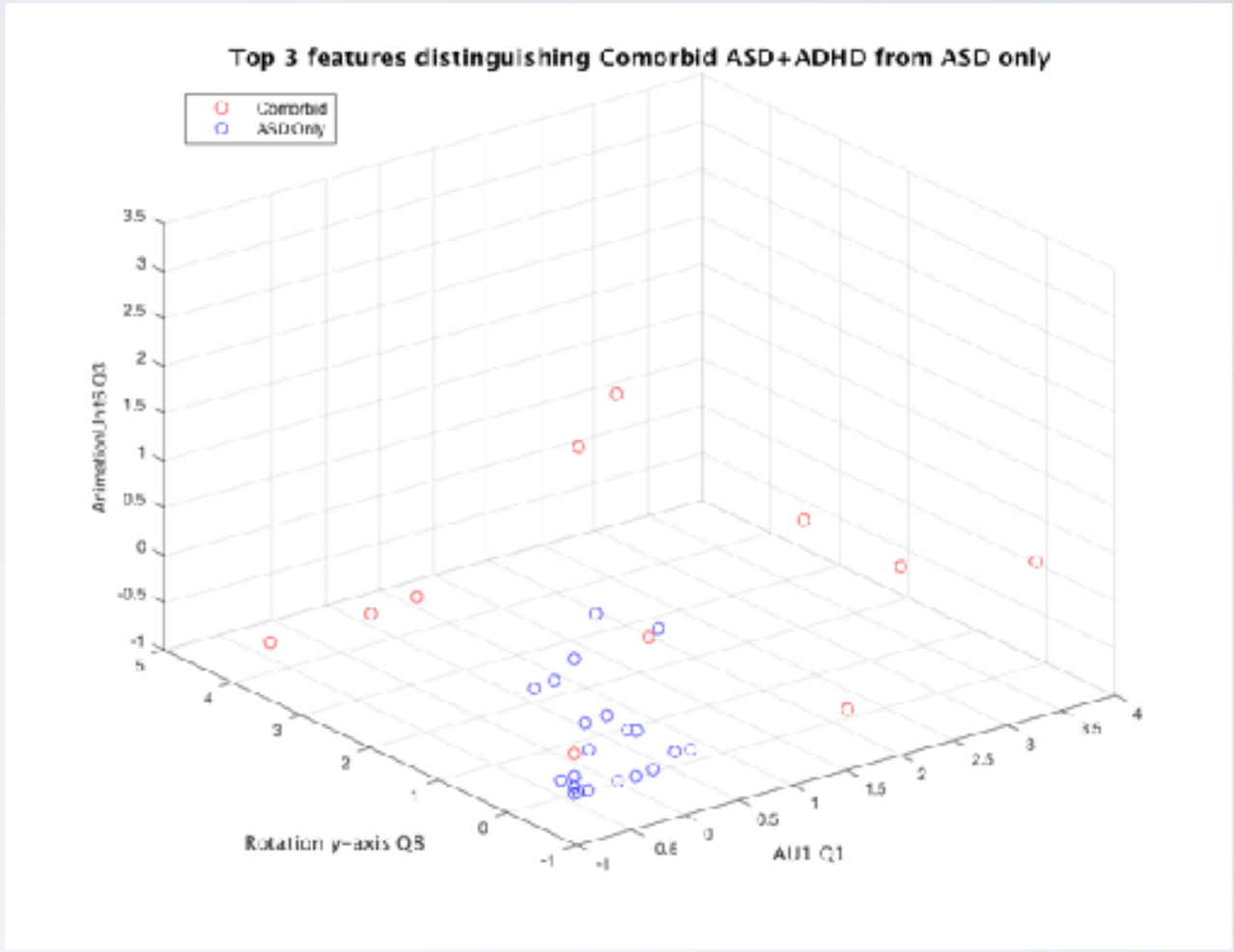
SELECTED FEATURES



- Top 30 features for classification of Controls vs Condition group. Each feature is represented by its feature type followed by the video segment number.
- E.g. AU15-S1 means that the feature corresponds to AU15 intensity histogram computed from the video segment corresponding to story 1 of the 'Strange stories' task.

COMORBID ASD+ADHD VS ASD

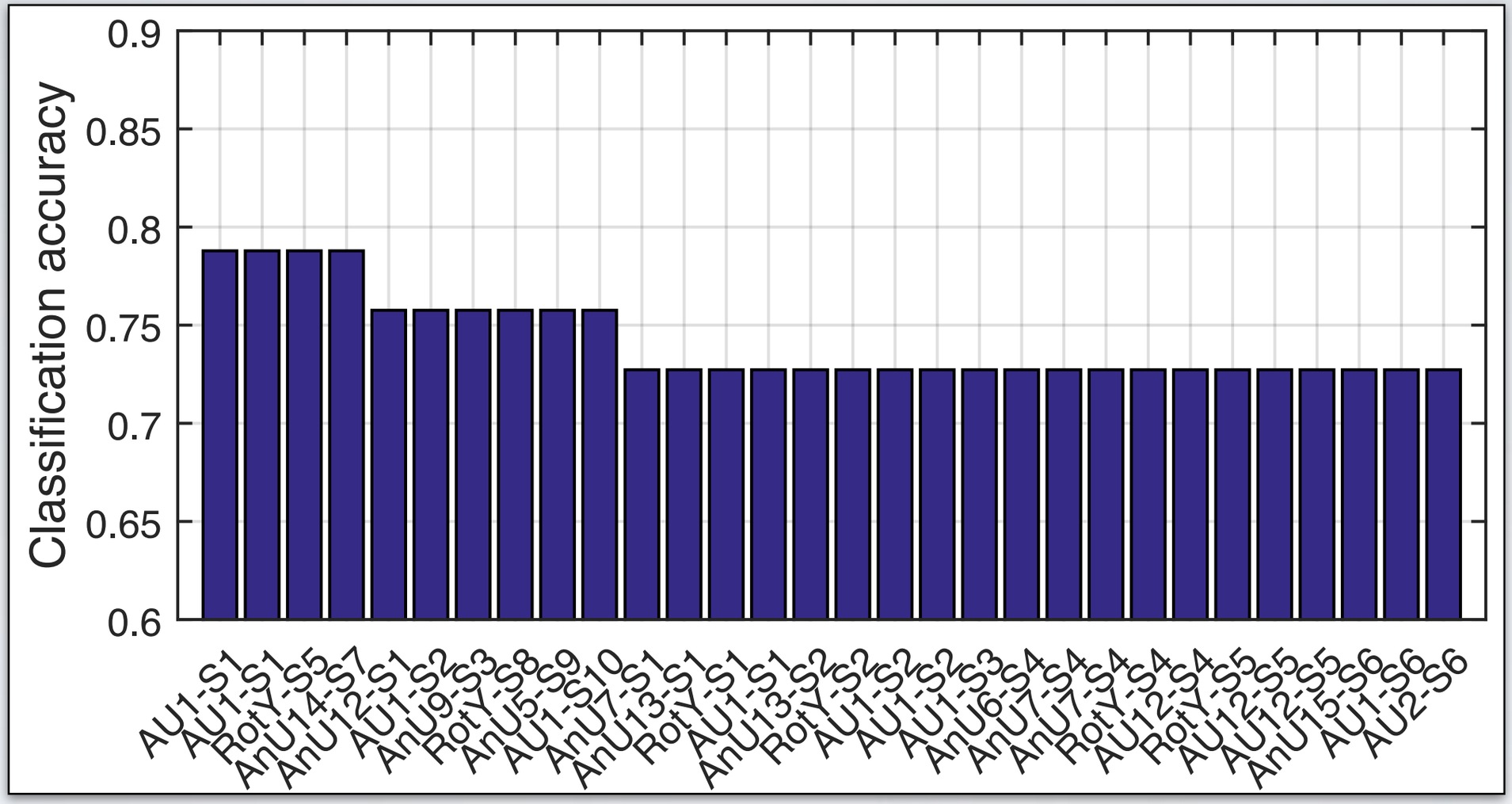
	Correct	Incorrect
Comorbid	9	2
ASD only	22	0



- Forward feature selection applied
- Classification using Support Vector Machines with a Radial Basis Function kernel
- Classification rate is 93.9% correct! Only two errors made in Leave-One-Out cross-validation
- Only four features selected, solution has only 9 support vectors (MDL= 36)

SELECTED FEATURES

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA



- Top 30 features for classification of Comorbid vs ASD Only group. Each feature is represented by its feature type followed by the video segment number.
- E.g. AU15-S1 means that the feature corresponds to AU15 intensity histogram computed from the video segment corresponding to story 1 of the 'Strange stories' task.

S. Jaiswal and M.F. Valstar, "*???*", proc. Int'l Conf Face and Gesture Recognition, 2017

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

DEPRESSION SEVERITY ESTIMATION

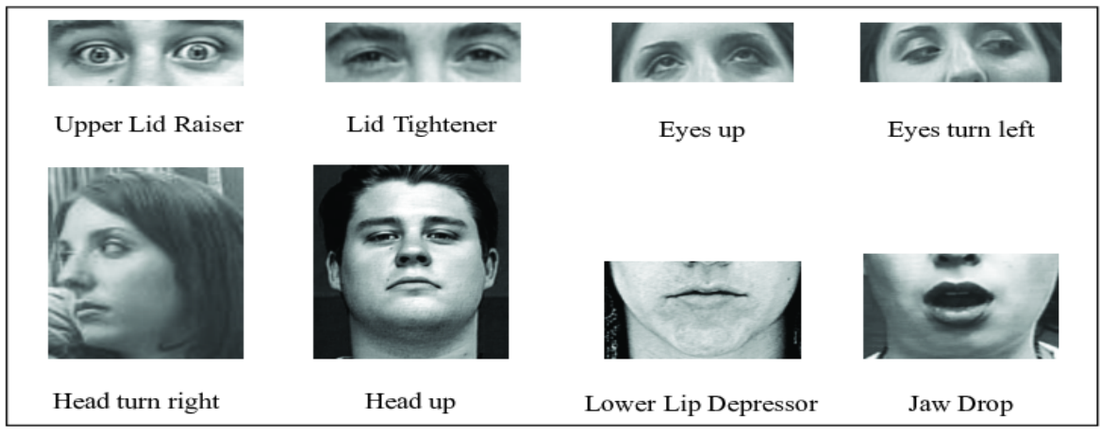
Current automatic approaches

- Using high dimensional audio/ video data that contain redundant information
- Predicting depression level based on a single frame or a small segment

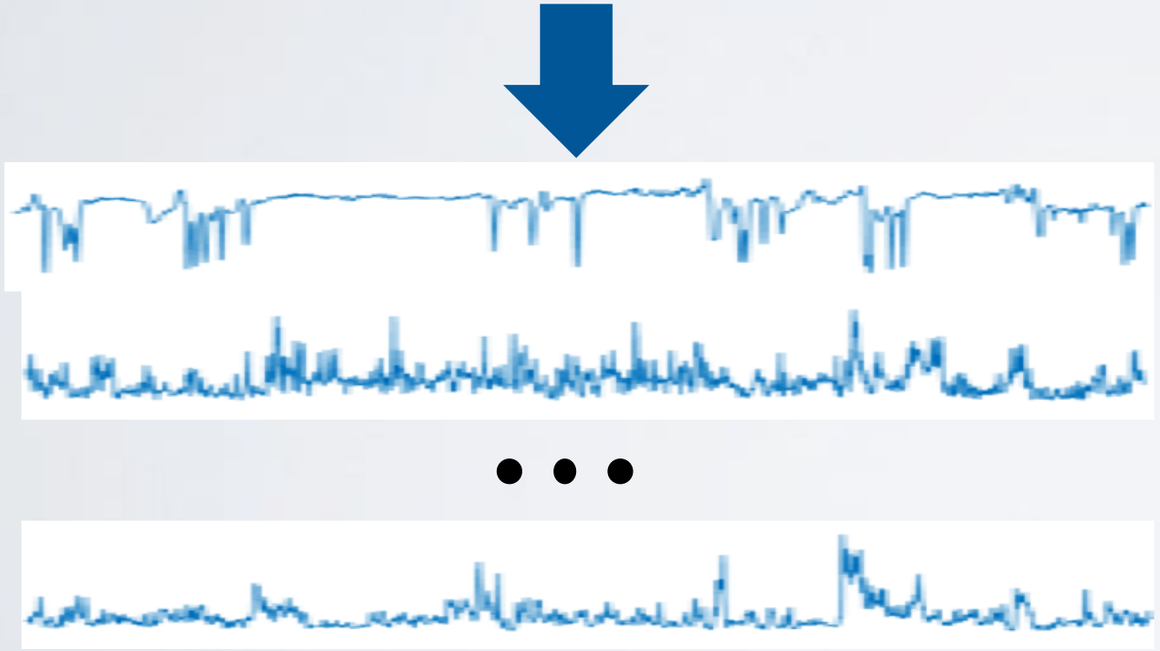
Our approach

- Using low dimensional human behaviour primitives as the input
- Predicting depression level based on a whole video

USING BEHAVIOUR PRIMITIVES



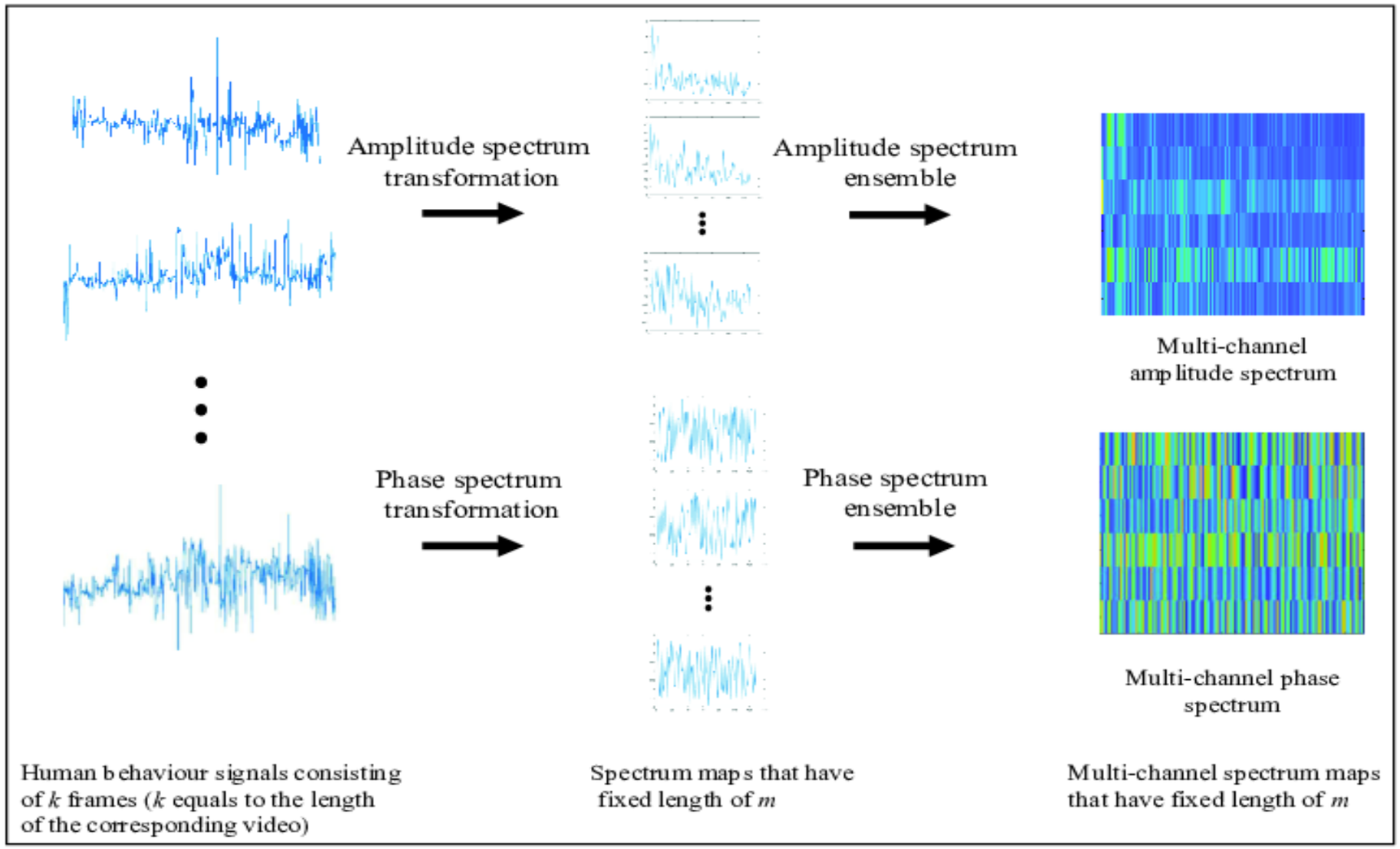
$$K \times N \times M \Rightarrow K \times S$$



If the resolution is 640×480 and $S = 10$, then

$$\frac{10}{640 \times 480} = \frac{1}{30720}$$

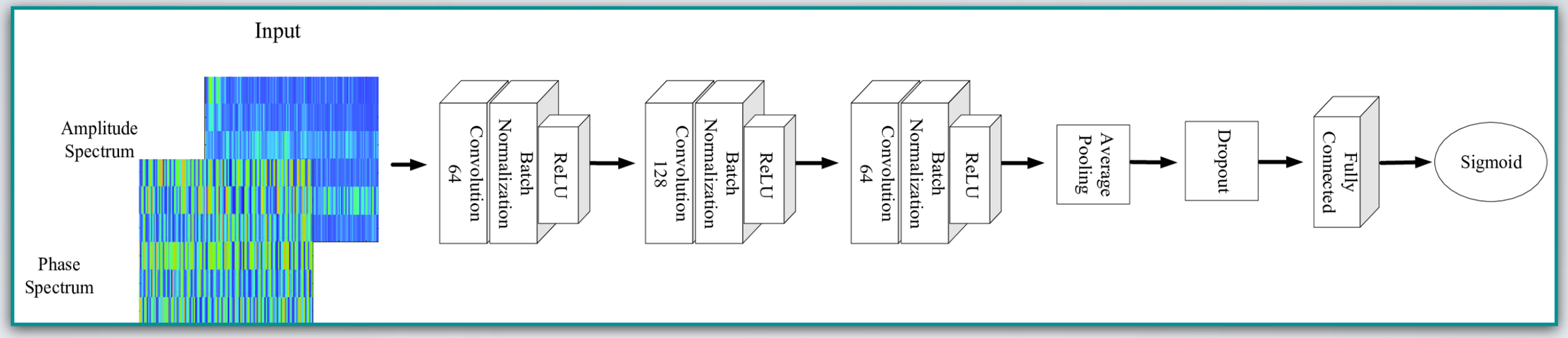
SPECTRAL CNN



Convert time-series signals to amplitude and phase spectrum maps that contains all and undistorted amplitude and temporal information.

SPECTRAL CNN

Take spectrum maps of all behaviour primitive signals as the multi-channel 1-D data and learn the correlations at the convolution layer level.



Method	RMSE
Nasir et al.	7.86
AVEC Baseline	7.13
Williamson et al.	6.45
Spectral CNN	**5.32**

Evaluation on the AVEC 2016 depression severity estimation challenge. Target range is [0, 16]

S. Song, L. Shen and M.F. Valstar (2018). *"Joint Action Unit localisation and intensity estimation through heatmap regression"*,

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

THE COMPUTATIONAL FACE

- To be published with Cambridge University Press
- Expected August 2019
- Approximately 300 pages

Part I: Underlying Principles Primer

1. Machine Learning Primer
2. Evaluating Hypotheses
3. Computer Vision Primer
4. Facial Anatomy
5. Face as Social Interface

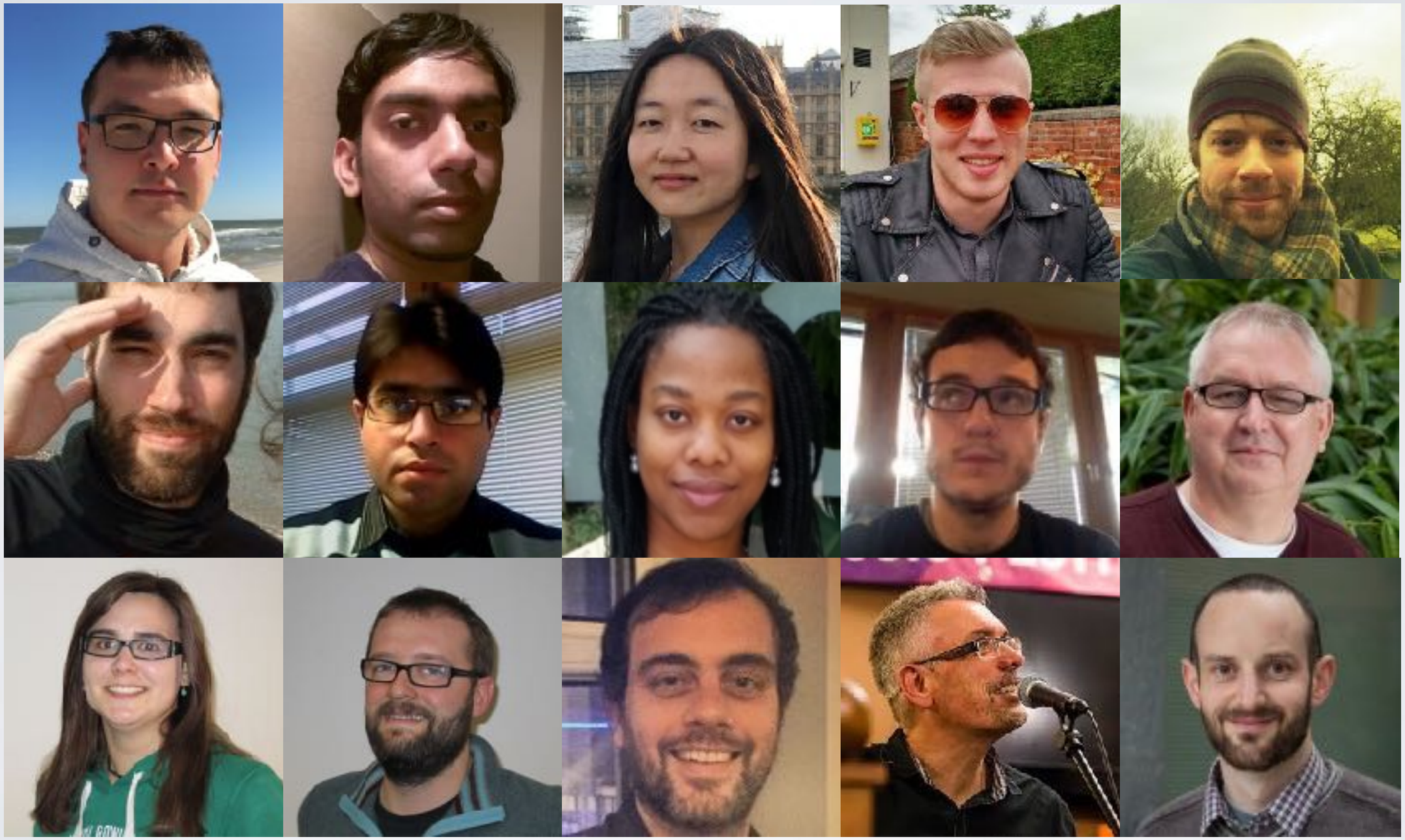
Part I: Theory of Automatic Face Analysis

6. Classical Analysis Pipeline
7. Computational Representation of the Face
8. Face Detection
9. Face Alignment & Tracking
10. Facial Expression Recognition
11. Face Recognition
12. Age and gender estimation
13. Personality estimation

Part III: Developing Practical Face Analysis Systems

14. Presenting eMax
15. Face Databases and Other Useful Resources
16. Face as Computational Interface
17. Face transfer and animation
18. Behaviomedics
19. Face analysis for market research
20. Facing Virtual Humans

TEAMWORK!



PRACTICAL WORK

Things to Look Out For

- Normalisation:
  - Head pose (Face registration)
  - Inter-personal differences
  - Lighting
  - Frame rate
- Occlusion

Design Choices

- Feature type:
  - Appearance
  - Geometric
- Classifiers
  - Static
  - Dynamic
- Resource consumption
  - Time
  - Memory

Learning and Evaluation

- Ensure Subject-independence
  - Also during parameter optimisation
- Often little data - use cross-validation
- Usually unbalanced data
  - Use AUC or 2AFC scores
  - NEVER use classification rate

Databases

Publicly available:

- MMI-Facial Expression DB
- Cohn-Kanade(+)
- SEMAINE
- McMaster PAIN database
- DISFA
- BP4D

On To Work!

- Download the code from:

  http://www.cs.nott.ac.uk/~pszmv/files/package.zip

- If the Zhu and Ramanan face detector and face alignment doesn't work, use iCCR which you can get from:

  https://github.com/ESanchezLozano/iCCR

  Or:

  http://www.cs.nott.ac.uk/~psxsj3/wacv_code.php

- Build a smile detector
- Follow the handout